Design and Implementation of an LLM-based Framework for Situated Workplace Support in the Interverse Office

Ayaka Takayanagi[1*], Satoki Ogiso, Mai Otsuki and Ryosuke Ichikari[1*]

[1] *National Institute of Advanced Industrial Science and Technology (AIST), Japan*

Abstract

The interverse office integrates physical workplaces, virtual offices, and digital twin data, integrating the value generated across both real and virtual environments into daily operations. To realize this environment, we accumulate a wealth of behavioral records, including user location and movement history. However, we have not yet been able to implement a system that appropriately translates context-dependent and ambiguous user needs into data acquisition and behavioral indicators. To address this challenge, we propose a context-aware LLM-based framework for the interverse office. This framework uses LLM to interpret natural language requests, references SQL for relevant behavioral indicators, and returns interpretable responses based on explicit computational procedures. Using this system, we evaluated 15 prompt-based tasks that ranked user IDs based on computational results. The results showed that 11 out of 15 tasks were correct or partially correct, suggesting that the proposed framework can partially support behavioral change in the office environment. At the same time, the results included some errors and delays, indicating the need for a more robust framework foundation in the future.

Keywords

interverse office, workplace support, large language model, MCP tools, situated interaction, behavioral data, digital twin, semantic spatial grounding

1. Introduction

In recent years, virtual work environments such as teleworking have expanded rapidly and have become an indispensable social infrastructure for work. Currently, physical and virtual offices are often separate, but if these can be integrated into a single environment, leading to increased productivity such as improved teamwork and estimation of worker states, it could lead to significant social transformation. Previous research defines interverse as a general term for technologies that return value co-created in the metaverse to the real universe rather than keeping it confined to the virtual environment [1,2]. As illustrated in Figure 1, the interverse office accumulates and visualizes behavioral data, including positioning information, from both the real universe and the virtual office. The important value of the interverse office lies not simply in implementing a metaverse office, but in integrating physical, virtual, and behavioral data across real and virtual environments to support more efficient and satisfying work practices.

However, this potential has not yet been fully realized. In today's interverse office environment, the importance of factors such as location, people's behavioral history, and current circumstances is not reflected in the user experience in a way that meets users' needs in the moment. Although behavioral data from both physical and virtual offices is being adequately recorded, people's needs and circumstances change constantly, making it extremely difficult to search for, extract, and integrate this behavioral data on a case-by-case basis to meet those needs. As a result, the wealth of data available in the interverse office is not being fully utilized to directly support workplace

APMAR'26: The 18th Asia-Pacific Workshop on Mixed and Augmented Reality, Aug. 3-5, 2026, Taipei, Taiwan

[1*] Corresponding authors.

a.takayanagi@aist.go.jp (A. Takayanagi); s.ogiso@aist.go.jp (S. Ogiso); mai.otsuki@aist.go.jp (M. Otsuki); r.ichikari@aist.go.jp (R. Ichikari)

0009-0007-3295-3235 (A. Takayanagi); 0000-0003-1338-3795 (S. Ogiso); 0000-0001-8303-465X (M. Otsuki); 0000-0002-1784-0538 (R. Ichikari)

behavior. Specifically, there is a challenge: Deriving appropriate search codes from ambiguous instructions to formulate suitable behavioral metrics and generate statistical results in the database; This unresolved issue serves as the motivation for the case study examined in this paper.

Interverse office of AIST(Kashiwa office)

Figure 1: An interverse office created by AIST. Both virtual and real-world positioning information of workers is accumulated. The image on the right visualizes real-world location information.

To realize these possibilities, we built a system that combines the office measurement data accumulated by our team with the Large Language Model (LLM) and Model Context Protocol (MCP) tools to orchestrate the entire process. In the current prototype design, behavioral data and workplace records are stored as structured data that can be searched via an SQL-like interface. Meanwhile, LLM interprets user requests, formulates appropriate metrics for those requests, searches the database for behavioral history, and returns responses translated into natural language based on statistical information. Therefore, we argue that by integrating behavioral data with LLM-based prompt interpretation, we can resolve the challenge and provide a practical interaction layer for the interverse office.

At this stage, addressing the challenge described above may create three possibilities for the interverse office. First, the interverse office is characterized by the fact that people can exist across both real and virtual environments. If the fragmentation described above can be reduced, collaboration may become easier even when people are spatially distributed. Second, solving these challenges may increase users' incentives to engage with the interverse. By using an AI-collaborative office, users can understand where activity is happening, who may be available, which places are suitable for a given purpose, and how physical and virtual resources should be coordinated. As a result, AI-collaborative user experiences could become possible within the office, potentially motivating users to engage more actively with the virtual office. Third, by combining historical data accumulated by the team with real-time behavioral data, the interverse office may eventually support proactive suggestions that encourage behavioral change, rather than only answering questions after they are asked. This study makes three contributions. First, to enhance the experiential value of the "Interverse," we construct a prototype capable of providing evidence-based responses to arbitrary user requests by utilizing behavioral data. Second, we implement a function that enables the immediate translation of ambiguous requests into SQL queries through the integration of LLMs and MCP tools. Third, we conduct a case study using existing data and analyze the results obtained from the LLM to evaluate the system's validity and consider directions for future improvement.

2. Background and Related Work

Research on interverse-related services emphasizes that value co-creation takes place across both the real and virtual worlds [1,2]. This distinction sets the interverse office apart from virtual offices, which merely replicate spaces such as offices within the metaverse. The advantage of using

interverse office is the ability to translate virtual data and insights back into the physical space; however, to achieve this, people, places, objects, and workplace resources must be recorded in both the physical and virtual worlds and processed appropriately according to the intended use. This challenge implies that AI-collaborative methods should be employed to help users understand and make decisions regarding, for example, where activities are taking place, how physical and virtual locations correspond, which resources are available, and how people can coordinate their actions in a distributed environment.

To realize AI-collaborative support in the interverse office, users' ambiguous natural-language requests must be translated into procedures that retrieve, interpret, and present workplace data. Natural-language database interfaces and text-to-SQL systems enable access to structured data without manual query writing, and recent surveys show the rapid development of LLM-based approaches [3]. Interactive explanation and repair systems, such as SQLucid and PleaSQLarify, further support users in understanding and correcting such queries [4,5]. Beyond database access, behavior retrieval and trajectory querying are relevant because interverse office support requires the interpretation of behavioral logs, location histories, room attributes, and activity records. Related studies include retrieval-enhanced behavior comprehension [6], natural-language visual querying of uncertain human trajectories [7], and LLM-based AIS trajectory querying [8]. In addition, Navigation Pixie shows how natural-language interaction can be connected to on-demand spatial assistance in a commercial metaverse [9].

Building on the interverse office perspective and the techniques reviewed above, our research aims to make physical and virtual workplace data usable in everyday work situations. In the interverse office, the key question is not only how to retrieve data, but also how to help users turn distributed information about people, places, activities, and resources into situated decisions. This requires context-dependent data extraction, translation of natural-language requests into operational procedures, and interpretable explanations of computational results. Previous research on text-to-SQL conversion, behavior retrieval, trajectory querying, and LLM-based navigation agents offer useful perspectives for achieving this goal. In this study, we use MCP tools to interpret natural language, retrieve request-relevant metrics from the database, and present numerical results in a human-readable format. By using these orchestration tools, we aim to design a framework in which LLM returns interpretable support for decision-making in the interverse office.

3. Framework Overview

This section introduces the conceptual workflow of the proposed interverse office support Framework. The purpose of this section is not to describe the architecture, but rather to conceptually explain how data-driven responses are generated in response to contextual user inquiries. These frameworks consist of three components: an LLM (Language Licensing Manager) that interacts with the user, a database that stores interverse office data, and MCP (Multi-Care Programming) tools that query the data based on instructions from the LLM (these tools are pre-regulated to retrieve relevant values from the database for specific purposes).

Figure 2 summarizes this conceptual workflow. When a user asks a question in natural language, the LLM breaks down the request into semantic elements such as the target person, time of day, and behavioral metrics to be referenced. Instead of creating a separate predefined function for every request, the LLM selects several MCP tools best suited to generating the metrics. The MCP tools query the behavioral data stored in the interverse office and pass it to the LLM. The interverse office data is organized as information including physical office information, virtual office information, a person's behavioral history across both, and workplace resource information. Based on this data, the LLM combines the results output by each MCP to provide the result to the user. Thus, the final response is an evidence-based response that includes not only the acquired values but also the

selected evidence and can be said to be a UX that helps users practically utilize the accumulated interverse office data.

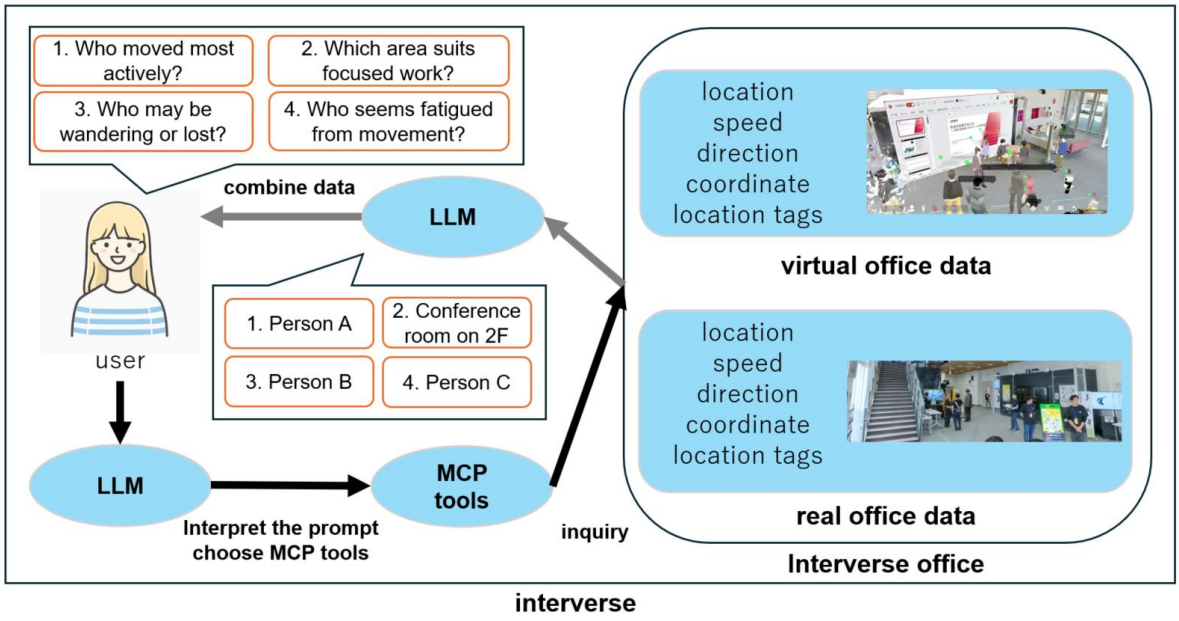

Figure 2: Conceptual workflow for responding to open-ended requests in the interverse office. The LLM system returns an evidence-based workplace support response.

4. Prototype Implementation

The prototype is organized as a set of grounding tools controlled by LLM. The behavioral-data tool returns position histories, dwell times, movement features, and occupancy statistics. A reasoning layer integrates these outputs to answer user requests. The implementation is regulated through MCP server so that LLM must use explicit tools rather than hallucinating places, people, or metrics.

In the prototype implementation, we used Meta's Llama 3.1 as the large language model and executed it in a local Ollama environment on a Windows 11 PC. The Ollama version used in this implementation was 0.30.7. The system implemented MCP tools as controlled interfaces between LLM and external data-processing functions. The behavioral data were stored in a PostgreSQL database, and database access was encapsulated inside predefined MCP tools rather than being freely generated by LLM. Each tool issued fixed SQL queries to retrieve user IDs, timestamps, positions, orientations, and derived movement features required for the corresponding calculation. This design allowed LLM to interpret user requests and select tools while keeping numerical computation and database access reproducible and constrained.

In our implementation, each MCP tool was implemented as a parameterized data-processing function rather than as a dynamically generated query. The tool contained the code for issuing the necessary SQL query and extracting the target numerical values from the PostgreSQL database, while variable conditions such as the target date, time range, and number of users were left as input parameters. When a user specified these conditions in a natural-language prompt, LLM mapped them to the corresponding tool arguments and invoked the appropriate MCP tool. The SQL structure and calculation logic therefore remained fixed inside the tool, whereas only the permitted parameters were filled according to the user request.

Tool use was regulated by the MCP server, which managed the available tool definitions, required arguments, permissible workflows, and prompt-level instructions. LLM was instructed not to create new tools, not to infer missing elements without asking follow-up questions, and to calculate

requested indicators only by combining the existing MCP tools. Before executing a calculation, LLM was required to present the intended calculation procedure to the user and confirm whether the procedure was acceptable. This regulation was implemented through workflow nodes for model invocation, tool selection, tool combination, and result generation, in which the permitted operations were strictly defined. As a result, the prototype supported flexible natural-language specification of dates and participants while maintaining constrained, reproducible access to behavioral data. The LLM's role was limited to interpreting the user request, selecting and combining available tools, requesting confirmation when necessary, and explaining the results, while SQL access and numerical computation remained inside predefined tool calls.

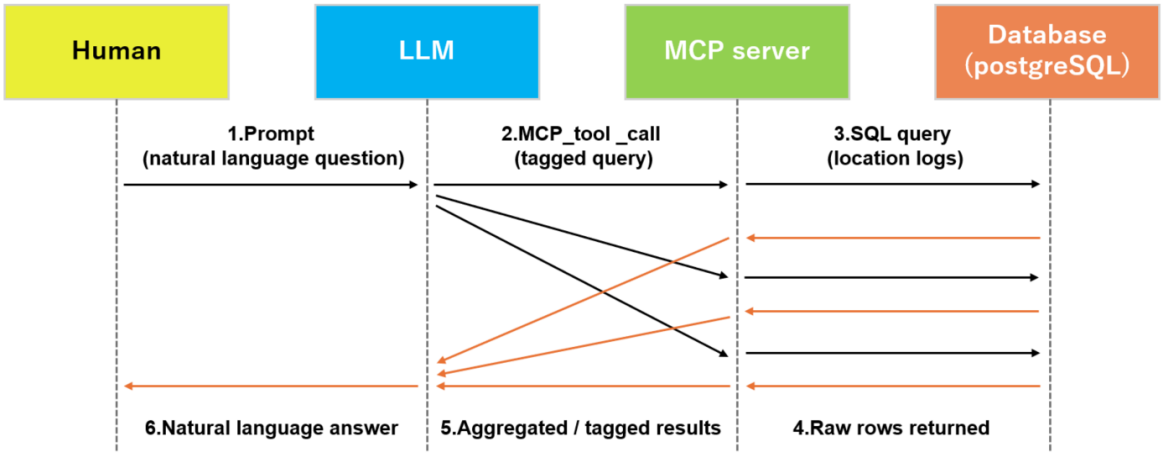

Figure 3: A diagram that illustrates data flow and prompts involved in LLM. It shows how the MCP tool returns metrics in response to any given prompt.

Figure 3 shows the regulated workflow used in the prototype. First, the user request is interpreted by LLM, the necessary behavioral metrics are identified, and it is determined whether existing MCP tools are sufficient and how to combine the MCP tools. If the request is not sufficiently specified, LLM asks additional questions. Otherwise, it presents the calculation procedure to the user and asks for confirmation. After confirmation, the selected MCP tool retrieves the necessary values through fixed SQL queries within the tool implementation, and LLM calculates the results of the MCP tool and outputs numerical results and rankings in natural language. For example, in response to the question, "Who looks most tired?", LLM first creates an index by combining MCP tools (for example, using positive and larger coefficients for walking distance and speed, as longer distances and speeds indicate greater fatigue). Once user permission is obtained, this index is accessed and the corresponding values for each MCP tool are queried. These values are then combined and calculated, translated into natural language, and displayed to the user by LLM.

5. Evaluation

To evaluate the formulation of behavioral metrics and the combination of tools, we designed a natural language prompt instructing the LLM to calculate mobility-related metrics-specifically walking speed (WS), walking distance (WD), and the number of turns (NT)based on existing indicators, and to sort the results by ID. The evaluation utilized location and orientation data from 13 employees who walked freely within the office while wearing beacons between 9:00 AM and 6:00 PM on November 3, 2025. Three pre-defined MCP tools were employed for this assessment: "avg moving speed summary" (calculating average walking speed), "distance summary" (calculating total walking distance), and "turn count summary" (calculating the number of sharp turns, defined as a change in orientation of 45 degrees or more). For any given prompt, the LLM was required to determine on a per-question basis-which MCP tools to use and how to weigh the metrics (WS, WD, NT) when formulating the calculation. Table 1 presents the calculation results generated by LLM.

Table 1
Example prompts for evaluating behavioral metric formulation.

No.	Prompt (natural language)	The formula provided by LLM	LLM's answer	Correct-ness
1	Which ID is most wandering around or lost?	-0.2* WD-0.3*WS+0.5*NT	Same as left	correct
2	Which ID seems the most fatigued based on their movement behavior?	0.1*WD+0.2*WS-0.3*NT	Sign inversion	incorrect
3	Which ID stayed in one place for too long and may have been inactive or waiting?	(1- WD) *100+(1-WS) *50	Too much (Meaning less)	incorrect
4	Which ID moved most efficiently between areas without unnecessary detours?	-0.5*WD+1.0*WS	Ranking changes	incorrect
5	Which ID showed the most inefficient movement pattern, such as repeated back-and-forth movement?	-0.2*WD+ 0.3*WS+0.5*NT	Same as left	correct
6	Which ID explored the widest range of areas during the day?	WD/WS	Same as left	correct
7	Which ID walked the longest distance?	WD	Normalized	partially correct
8	Which ID walked the fastest on average?	WS	Meaning less	incorrect
9	Which ID changed direction most frequently?	NT	Some IDs missing	partially correct
10	Which ID seemed to be moving around the most actively?	0.5*WD+0.5* WS	Same as left	correct
11	Which ID moved most efficiently, with long walking distance and fewer turns?	0.6*WD-0.4*NT	Same as left	correct
12	Which ID showed signs of hesitation while moving, based on frequent turns and low walking speed?	0.6*NT-0.4* WS	Some IDs missing	partially correct
13	Which ID seemed to be in a hurry, based on high walking speed and long walking distance?	0.6*WD+0.4*WS	Some IDs missing	partially correct
14	Which ID seemed to move with the least effort, based on short distance, low speed, and few turns?	-0.3*WD-0.2*WS-0.5*NT	Same as left	correct
15	Which ID had the highest movement intensity, considering walking speed, walking distance, and number of turns together?	0.3*WD+0.4*WS-0.1*NT	Some IDs missing	partially correct

In the prompt design, LLM was controlled to combine MCP tools according to any user instructions, define the necessary behavioral metrics, and calculate them only after receiving confirmation from the user. The prompt also instructed LLM to output both numerical results and corresponding ID rankings.

5.1. Correct answer

In the following three subsections, we compare the metrics calculated by the LLM against those calculated by humans for each question, based on the SQL data. An output is considered correct if all requested IDs and metrics are presented as instructed and the rankings and numerical values are accurate. Tables 2 through 7 summarize the responses to the six questions where the LLM's answers matched the results of the manual verification. The metrics used are those from the "The formula provided by LLM" column in Table 1. In these tables, the score is not a physical unit such as meters or seconds, but a prompt-specific composite index calculated from the behavioral indicators selected by the LLM. Unless otherwise noted, a higher score means that the ID more strongly matches the behavioral interpretation encoded by the formula for that question. Negative values can occur when the formula includes negatively weighted normalized variables; therefore, they should be interpreted

as relative positions within the prompt-specific score range rather than as directly negative real-world quantities.

Table 2

This is a comparison of rankings-determined by an LLM versus a human expert-regarding the "wandering prompt" in Question 1. The "Rank" column indicates positions from 1st to 13th. The "LLM" column lists the IDs and scores for each rank as generated by the LLM, while the "Human Expert" column lists the IDs and scores calculated by a human using SQL data-following the formula output by the LLM-which served as the benchmark for correctness. This same table format is used for the subsequent comparison tables as well.

Question 1: Wandering: Which ID is wandering around or lost?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	293, 292, 284, 290, 287, 286, 282, 289, 281, 291, 288, 285, 283	0.405483, 0.263923, 0.131744, 0.087666, 0.041191, 0.035015, 0.019096, 0.008621, 0.008609, -0.091430, -0.124552, -0.271324, -0.300099	293, 292, 284, 290, 287, 286, 282, 289, 281, 291, 288, 285, 283	0.405483, 0.263923, 0.131744, 0.087666, 0.041191, 0.035015, 0.019096, 0.008621, 0.008609, -0.091430, -0.124552, -0.271324, -0.300099

Table 2 can be read as a concrete example of this comparison format. ID 293 is ranked first with a score of 0.405483, indicating that, under the prompt-specific wandering index, this ID showed the strongest relative match to the behavioral interpretation of wandering or being lost. In contrast, ID 283 is ranked last with a score of −0.300099, indicating the weakest match to this prompt-specific interpretation. Because the ranked IDs and scores are identical between the LLM and Human Expert columns, this output was judged correct. The formula used for Table 2 was −0.2×WD −0.3×WS +0.5×NT, as shown in the "The formula provided by LLM" column in Table 1. In this calculation, WD, WS, and NT were normalized using $(x - min)/(max - min)$, so each normalized variable was constrained to the range from 0 to 1. As a result, the total possible score range was bounded between −0.5 and 0.5, which explains why the scores in Table 2 fall within this interval.

Table 3

Comparison of LLM and human-expert rankings for the inefficient movement prompt in Question 5.

Question 5: Inefficient Movement: Which ID showed the most inefficient movement pattern?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	292, 293, 283, 285, 286, 284, 290, 288, 287, 291, 282, 289, 281	missing	292, 293, 283, 285, 286, 284, 290, 288, 287, 291, 282, 289, 281	0.580418, 0.498772, 0.299901, 0.270191, 0.260573, 0.249274, 0.231639, 0.158978, 0.151753, 0.087981, 0.048580, 0.008621, 0.008609

For Table 3, the score is calculated as −0.2×WD +0.3×WS +0.5×NT. With WD, WS, and NT normalized to the range from 0 to 1, the possible score range is bounded between −0.2 and 0.8.

Table 4

Comparison of LLM and human-expert rankings for area coverage prompt in Question 6.

Question 6: Area Coverage: Which ID explored the widest range of areas during the day?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	281, 289, 291, 284, 287, 286, 293, 290, 292, 282, 285, 283, 288	missing	291, 284, 287, 286, 293, 290, 292, 282, 285, 283, 288, 289, 281	3.344277, 2.838791, 2.762368, 2.620623, 1.539502, 0.804814, 0.084022, 0.065617, 0.003141, 0.000495, 0.000297, NULL, NULL
Zero-speed cases are treated as NULL in SQL.				

For Table 4, the score is calculated as the ratio WD/WS. Unlike the weighted normalized scores, this ratio is not bound between 0 and 1; it becomes larger when walking distance is large relative to average walking speed. Cases with zero speed are treated as NULL to avoid division by zero.

Table 5

Comparison of LLM and human-expert rankings for the activity level prompt in Question 10.

Question 10: Activity Level: Which ID seemed to be moving around the most actively?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	286, 291, 283, 285, 284, 287, 292, 288, 290, 293, 282, 281, 289	missing	286, 291, 283, 285, 284, 287, 292, 288, 290, 293, 282, 281, 289	0.680550, 0.649509, 0.500248, 0.452680, 0.375980, 0.346645, 0.285906, 0.236345, 0.216537, 0.197422, 0.026182, 0.000029, 0

For Table 5, the score is calculated as 0.5×WD + 0.5×WS. Because WD and WS are normalized to the range from 0 to 1, the resulting activity-level score is also bounded between 0 and 1.

Table 6

Comparison of LLM and human-expert rankings for route efficiency prompt in Question 11.

Question 11: Route Efficiency: Which ID moved most efficiently, with long walking distance and fewer turns?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	291, 286, 287, 284, 285, 283, 281, 289, 288, 282, 290, 293, 292	0.441379, 0.315240, 0.146791, 0.092266, 0.001701, 0.000297, -0.006862, -0.006897, -0.013709, -0.025652, -0.042749, -0.256382, -0.318235	291, 286, 287, 284, 285, 283, 281, 289, 288, 282, 290, 293, 292	0.441379, 0.315240, 0.146791, 0.092266, 0.001701, 0.000297, -0.006862, -0.006897, -0.013709, -0.025652, -0.042749, -0.256382, -0.318235

For Table 6, the score is calculated as 0.6×WD − 0.4×NT. With WD and NT normalized to the range from 0 to 1, the possible score range is bounded between −0.4 and 0.6.

Table 7
Comparison of LLM and human-expert rankings for the low effort movement prompt in Question 14.

Question 14: Low Effort Movement: Which ID seemed to move with the least effort?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	289, 281, 282, 288, 285, 283, 290, 287, 284, 292, 291, 293, 286	-0.008621, -0.008638, -0.045278, -0.111793, -0.181355, -0.200149, -0.304203, -0.387836, -0.507724, -0.549829, -0.558080, -0.602905, -0.715564	289, 281, 282, 288, 285, 283, 290, 287, 284, 292, 291, 293, 286	-0.008621, -0.008638, -0.045278, -0.111793, -0.181355, -0.200149, -0.304203, -0.387836, -0.507724, -0.549829, -0.558080, -0.602905, -0.715564

For Table 7, the score is calculated as $-0.3\times WD\ -0.2\times WS\ -0.5\times NT$. Because all three normalized variables contribute negatively, the possible score range is bounded between −1.0 and 0; values closer to 0 indicate a stronger match to the low-effort movement interpretation.

5.2. Partially correct answer

Partially correct output includes cases where the rankings are not sorted in descending order, but the IDs and numerical values are matched. Also, even if some rankings are missing, the output will be categorized as partially correct if the numerical values in the displayed portion are accurate. Tables 8 through 12 summarize the answers to five questions where the LLM and manual check results partially matched.

Table 8
Comparison of LLM and human-expert rankings for the walk distance prompt in Question 7.

Question 7: Walk Distance: Which ID walked the longest distance?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	282, 287, 292, 286, 284, 288, 291, 285, 293, 289, 281, 283, missing	0.003224, 0.509020, 0.044321, 0.985170, 0.556075, 0.000140, 1.000000, 0.002834, 0.239363, 0.000057, 0.000495, missing	291, 286, 284, 287, 293, 290, 292, 282, 285, 283, 288, 281, 289	7937.348, 7819.646, 4414.102, 4040.645, 1900.487, 1533.469, 352.520, 26.352, 23.257, 4.694, 1.874, 1.217, 0.762
LLM returned normalized values and omitted ID 290; missing is marked for the absent ranking/score. When sorted in descending order, it matches the SQL sorting order.				

For Table 8, the intended metric is WD, which represents the raw walking distance derived from the SQL data. Unlike the composite indices in Tables 2–7, WD is used here as an exception in its original, non-normalized scale; therefore, the human-expert column reports the correct walking-distance values. The LLM output was partially correct rather than fully correct because it retrieved most of the relevant IDs but incorrectly reported normalized scores instead of the requested raw WD values and omitted ID 290. In addition, the LLM output itself was not presented in descending order. However, if the LLM's normalized WD values are sorted in descending order, the resulting ranking is consistent with the human-expert ranking for the displayed IDs. Thus, the error in Table 8 was a

combination of inappropriate normalization, incorrect output ordering, and incomplete output, while the underlying distance-based ranking was still partly preserved.

Table 9
Comparison of LLM and human-expert rankings for the direction changes prompt in Question 9.

Question 9: Direction Changes: Which ID changed direction most frequently?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	293, 292, 286, 284, 291, 287, 290, 282, 281, 289, missing, missing, missing	missing	293, 292, 286, 284, 287/290/291, 287/290/291, 287/290/291, 282, 288, 289/281, 289/281, 283/285, 283/285	58, 50, 40, 35, 23, 23, 23, 4, 2, 1, 1, 0, 0
LLM omitted ID 288 and zero-turn IDs; missing is marked for omitted rows.				

For Table 9, the intended metric is NT, which represents the raw turn count derived from the SQL data. Because NT is used here as a non-normalized count, it is not bounded between 0 and 1; the minimum possible value is 0, and the upper bound depends on the number of direction changes observed in the data. Therefore, the human-expert column reports the correct raw turn-count values. The LLM output was partially correct because it identified several high-turn-count IDs in the correct order, but it omitted ID 288, the zero-turn IDs, and the corresponding scores.

Table 10
Comparison of LLM and human-expert rankings for the hesitation prompt in Question 12.

Question 12: Hesitation: Which ID showed signs of hesitation while moving?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	293, 292, 284, 286, 287, missing, missing, missing, missing, missing, missing, missing, missing	0.537808, 0.306245, 0.283715, 0.263421, 0.164223, missing, missing, missing, missing, missing, missing	293, 292, 284, 286, 287, 290, 291, 282, 289/281, 289/281, 288, 285, 283	0.537808, 0.306245, 0.283715, 0.263421, 0.164223, 0.141949, 0.118324, 0.021724, 0.010345, 0.010345, -0.168330, -0.361010, -0.400000
LLM omitted IDs from 6th place onwards.				

For Table 10, the score is calculated as 0.6×NT − 0.4×WS. Because NT and WS are normalized to the range from 0 to 1, the possible score range is bounded between −0.4 and 0.6. The LLM output was partially correct because the displayed IDs and scores from 1st to 5th place matched the human-expert results, but it omitted all IDs from 6th place onward. Therefore, the error in Table 10 was not a scale or formula error, but an output-completeness error in which the lower-ranked IDs and their scores were not presented.

Table 11
Comparison of LLM and human-expert rankings for the hurry prompt in Question 13.

Question 13: Hurry: Which ID seemed to be in a hurry?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	286, 291, 284, 283, 287, 285, 292, 290, 293, 288, missing, missing, missing	0.741474, 0.719607, 0.411999, 0.400297, 0.379120, 0.362711, 0.237589, 0.211854, 0.205810, 0.189104, missing, missing, missing	286, 291, 284, 283, 287, 285, 292, 290, 293, 288, 282, 281, 289	0.741474, 0.719607, 0.411999, 0.400297, 0.379120, 0.362711, 0.237589, 0.211854, 0.205810, 0.189104, 0.021590, 0.000034, 0
LLM omitted IDs from 11th place onwards.				

For Table 11, the score is calculated as 0.6×WD + 0.4×WS. With WD and WS normalized to the range from 0 to 1, the possible score range is bounded between 0 and 1. The LLM output was partially correct because the displayed IDs and scores from 1st to 10th place matched the human-expert results, but it omitted ID 282, 281, and 289 from 11th place onward. Thus, the error in Table 11 was an output-completeness error.

Table 12
Comparison of LLM and human-expert rankings for the movement intensity prompt in Question 15.

Question 15: Movement Intensity: Which ID had the highest movement intensity?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	missing, 291, 286, 285, 287, 288, 284, 292, 290, 293, 282, 281, 289	missing	283, 291, 286, 285, 287, 288, 284, 292, 290, 293, 282, 281, 289	0.400149, 0.379952, 0.376957, 0.361860, 0.186759, 0.185614, 0.184832, 0.138086, 0.114263, 0.034001, 0.013726, -0.001707, -0.001724
LLM omitted ID 283.				

For Table 12, the score is calculated as 0.3×WD + 0.4×WS − 0.1×NT. Because WD, WS, and NT are normalized to the range from 0 to 1, the possible score range is bounded between −0.1 and 0.7. The LLM output was partially correct because the ranking and IDs from 2nd place onward matched the human-expert results, but it omitted ID 283, which should have been ranked first with the highest score. Therefore, the error in Table 12 was a top-ranked ID omission.

5.3. Incorrect answer

Incorrect output included ranking errors such as two IDs being swapped, and output where the calculated result differed significantly from the expected answer. Tables 13 through 16 summarize the answers to the four questions where the LLM and manual check results did not match.

Table 13

Comparison of LLM and human-expert rankings for the fatigue prompt in Question 2.

Question 2: Fatigue: Which ID seems the most fatigued based on movement behavior?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	293, 292, 284, 290, 286, 287, 291, 288, 285, 283, 282, 289, 281	-0.244968, -0.148690, -0.086250, -0.051663, -0.033194, -0.031210, -0.040838, -0.084179, -0.180788, -0.200050, -0.010539, -0.005172, -0.005167	283, 285, 288, 291, 281, 289, 282, 287, 286, 290, 284, 292, 293	0.200050, 0.180788, 0.084179, 0.040838, -0.005167, -0.005172, -0.010539, -0.031210, -0.033194, -0.051663, -0.086250, -0.148690, -0.244968
The LLM ranking is reversed relative to SQL, and several ID (283, 285, 288, 291) signs differ.				

For Table 13, the score is calculated as 0.1×WD + 0.2×WS − 0.3×NT. Because WD, WS, and NT are normalized to the range from 0 to 1, the possible score range is bounded between −0.3 and 0.3. The human-expert column reports the reference ranking obtained by applying this formula to the normalized SQL-derived values. The LLM output was incorrect because the signs of the resulting scores were reversed for the leading fatigue-related IDs, and the ranking was therefore nearly inverted relative to the human-expert result. In other words, IDs that should have had higher fatigue scores were assigned lower or negative scores by the LLM.

Table 14

Comparison of LLM and human-expert rankings for the inactive ID prompt in Question 3.

Question 3: Inactive: Which ID stayed in one place for too long?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	281, 289, 291, 284, 287, 286, 293, 290, 292, 282, 285, 283, 288	1216818456.653, 761610559.397, 31282.858, 26556.499, 25841.989, 24513.726, 14405.114, 7531.368, 787.585, 631.994, 30.369, 5.531, 4.675	289, 281, 282, 288, 292, 290, 293, 285, 283, 287, 284, 291, 286	150.000, 149.994, 147.221, 126.358, 119.193, 118.690, 118.290, 104.590, 99.950, 89.885, 84.598, 35.049, 32.687
Meaningless error				

For Table 14, the score is calculated as (1−WD) × 100 + (1−WS) × 50. Because WD and WS are normalized to the range from 0 to 1, the possible score range is bounded between 0 and 150. The human-expert column reports the reference values calculated from the normalized SQL-derived data, and the resulting scores fall within this expected range. The LLM output was incorrect because it produced extremely large values far outside the valid range and a ranking that did not match the human-expert result. To investigate the source of the discrepancy, we also manually checked the values before normalization, but those raw values did not correspond to the LLM scores either. Therefore, the cause of the Table 14 error could not be interpreted from the available output.

Table 15
Comparison of LLM and human-expert rankings for the efficient movement prompt in Question 4.

Question 4: Efficient Movement: Which ID moved most efficiently between areas?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–13	283, 285, 288, 292, 290, 282, 293, 289, 281, 287, 284, 286, 291	missing	283, 285, 292, 288, 290, 282, 293, 289, 281, 287, 284, 286, 291	0.999752, 0.901107, 0.505330, 0.472480, 0.143395, 0.047527, 0.035799, 0, -0.000029, -0.070241, -0.082153, -0.116655, -0.200982
The rankings for ID 288 and ID 292 have been swapped.				

For Table 15, the score is calculated as −0.5×WD + 1.0×WS. Because WD and WS are normalized to the range from 0 to 1, the possible score range is bounded between −0.5 and 1.0. The LLM output was incorrect because the final ordering did not fully match the human-expert ranking: specifically, ID 288 and ID 292 were swapped. This means that the LLM identified most of the same ranked IDs but failed to preserve the exact descending order implied by the normalized scores. Therefore, the error in Table 15 was a ranking-order error rather than a scale error. As the LLM's output lacked the actual score values, it was impossible to determine whether the swapped IDs resulted from a numerical error or an error during the subsequent sorting process.

Table 16

Question 8: Average Speed: Which ID walked the fastest on average?				
Rank	LLM		Human Expert	
	Ranked IDs	Scores of the ranked IDs	Ranked IDs	Scores of the ranked IDs
1–11	283, 285, 286, 284, 293, 292, 287, 291, 288, 290, missing	missing	283, 285, 292, 288, 286, 291, 290, 284, 287, 293, 282	0.848538, 0.765826, 0.447596, 0.400977, 0.318991, 0.253728, 0.203611, 0.166216, 0.156360, 0.131931, 0.041696
There are significant differences from 3rd place onwards (meaningless).				

For Table 16, the intended metric is WS, which represents the raw average walking speed derived from the SQL data. Unlike the composite normalized indices, WS is used here in its original, non-normalized scale; therefore, it is not bounded between 0 and 1, and the human-expert column reports the correct average-speed values. The LLM output was incorrect because, although the first two ranked IDs matched the human-expert result, the ranking differed substantially from 3rd place onward, and ID 282 was omitted. In addition, the LLM did not output any score values, making it impossible to determine whether the discrepancy was caused by incorrect speed calculation, inappropriate normalization, omission of a retrieved value, or an error during ranking.

6. Discussion

6.1. Interpretation of the evaluation

Evaluation of the system revealed that 6 out of 15 case studies (40%) yielded correct answers, a figure that rose to 11 (73%) when including partially correct cases. These results suggest that the proposed framework can partially address the challenge outlined in the introduction: translating arbitrary and ambiguous queries regarding the workplace into actionable behavioral metrics. In the prompt-based evaluation, the LLM was able to interpret multiple natural language requests and formulate queries

by selecting or combining predefined MCP tools. For instance, prompts requiring only a single MCP tool-such as those for walking distance, average walking speed, or number of turns-were interpreted as direct mappings from user requests to specific metrics. Furthermore, for prompts requiring combinations of multiple metrics-such as "walking around and lost," inefficient movement, activity levels, route efficiency, and low-intensity movement-the system produced responses that appeared highly plausible, even in the absence of formal objective validation. This indicates that the LLM was not merely extracting existing SQL fields but was constructing higher-level workplace metrics by utilizing available behavioral features as evidence. In this sense, the proposed framework transcends conventional natural language database access methods like PleaSQLarify; it facilitates a form of "semantic grounding" [5] wherein ambiguous workplace concepts are transformed into measurable behavioral components through controlled tool usage. Additionally, based on the formulated queries, the system successfully utilized MCP tools to access SQL data, allowing the LLM to compute results and output rankings of metrics or IDs. This demonstrates that the system can function not only as a database access interface but also as an integration layer linking user intent with workplace behavioral logs within an "interverse office" environment. Such systems can help users understand their own and others' workplace activities and-by prompting the recognition of inefficient movement patterns or the more effective coordination of physical and virtual workplace resources-may ultimately contribute to behavioral change.

6.2. Error Types as Design Suggestions

In the evaluation of use cases, 4 out of 15 questions (27%) yielded incorrect results. Three primary types of errors emerged during the processes of formulating logic, referencing SQL, and performing calculations or generating responses via the LLM. The first type involved errors related to calculation or scaling, such as sign inversion or inappropriate normalization; Questions 2 and 7 are prime examples. In the former, the sign of the calculation result for a specific ID was reversed, while in the latter, the calculation proceeded without the necessary normalization. The second type concerned output completeness, where some IDs present in the source data were omitted. Question 12 exemplifies this issue, with the number of missing IDs varying by question. The third type involved errors in ranking or ordering; while the numerical values and IDs were generally retrieved correctly, the final sort order was not descending. Specifically, Question 2 resulted in an ascending order, and Question 7 produced a disordered sequence. Fourthly, there were errors with unidentified causes. For instance, in Question 3, the LLM outputs an absurdly large value—one that did not even correspond to the pre-normalization value of the correct answer. Additionally, in Question 4, where the rankings of certain IDs were swapped, the absence of the "score" in the output made it impossible to determine whether the issue lay in the LLM's calculation or in the final sorting process.

These errors indicate that the current prototype's primary weakness lies with LLM, particularly in the latter stages of the process. While the necessary values were often successfully retrieved via predefined MCP tools, errors frequently occurred during the LLM-driven phases of weighted calculation, normalization, ranking, or final output generation. This points to a crucial design requirement: although LLMs can flexibly formulate behavioral metrics based on ambiguous prompts, they should not be solely responsible for subsequent stages involving numerical calculation or ranking. Instead, these stages should be handled through more rigorous procedures within the tools themselves or via a dedicated validation layer. First, regarding errors related to calculations or scaling, the system must explicitly manage units, normalization methods, sign handling, valid ranges, and sorting order. Since comparing generated results against ground-truth data is difficult in actual operations, a viable strategy is to perform the same calculation multiple times in parallel and verify the consistency of the results. Second, errors concerning the display of calculated data should be managed via MCP tools; the MCP should verify formatting—such as whether the number of output IDs matches the number of target IDs retrieved from the database and whether the list is sorted in descending order—prior to final output. If the displayed content fails to meet these controls, the system must reject the response and regenerate the calculation or output. Finally, regarding errors

with unknown causes, it is necessary to log SQL return values and intermediate calculation steps to make the calculation process transparent.

Addressing errors in this manner increases the likelihood of achieving a highly reliable workplace support system. This requires a robust feedback system covering the calculation process, results, and output completeness. This implies that MCP tools should play a more critical role—acting not merely as interfaces for data retrieval, but as control components that constrain and validate LLM calculations and output formats.

6.3. Limitations and Future Directions

The current design faces several limitations. First, it relies on strict tool management; if tool usage is not properly controlled, the LLM may misinterpret spatial domains, user states, or actions. This challenge stems largely from the performance constraints of LLMs, particularly when running locally. We plan to address this by accessing higher-performance local LLMs via orchestration tools that ensure secure connections. Second, while the current framework responds to user queries by combining pre-defined MCP tools with structured behavioral data, it does not yet support open-vocabulary spatial grounding for arbitrary locations or objects. Moving forward, we plan to integrate Vision-Language Models (VLMs) into the system to combine positional and behavioral data, thereby expanding the scope of our responses to users.

7. Conclusion and Future Work

This paper proposes an LLM-based framework for context-aware workplace support in the interverse office. Our main contribution lies not in the implementation of a specific model, but in the design of a framework that links natural language requests with spatial context, behavioral logs, and workplace-related information. Through scenario-based case studies in the interverse office, we confirmed that this framework can output metrics as numerical values with a certain degree of accuracy for arbitrary queries such as fatigue level, dwell time, and walking efficiency. These results suggest that the interverse office will become more useful if users can ask context-aware questions and receive answers based on clear evidence and assumptions. In future research, we plan to incorporate visual understanding methods using open vocabulary as a component that will form the basis of semantic location search in the future. Then, we will evaluate these systems from various perspectives using scenarios with real-universe workplace data.

Acknowledgements

This work was supported by Council for Science, Technology and Innovation, “Cross-ministerial Strategic Innovation Promotion Program (SIP), Development of foundational technologies and rules for expansion of the virtual economy"（JPJ012495）. (funding agency: NEDO)

Declaration on Generative AI

During the preparation of this draft, the authors used an LLM-based writing assistant to organize ideas, draft text, and refine wording. The authors will review and edit the content as needed and take full responsibility for the final publication.

References

[1] K. Osawa, K. Watanabe, B. Q. Ho, T. Takka, Examining RRI process of interverse services: Focusing on an ELSI response of business operators, in: The 34th Annual RESER Conference, Helsinki, Finland, 2024.

[2] L. B. Q. Ho, K. Osawa, K. Watanabe, T. Okuma, Developing a servicescapes framework for the interverse: Promoting value co-creation across real and virtual worlds, in: The 34th Annual RESER Conference, Helsinki, Finland, 2024.

[3] L. Shi, Z. Tang, N. Zhang, X. Zhang, Z. Yang, A survey on employing large language models for text-to-SQL tasks, ACM Computing Surveys 58(2) (2025) 1-37. doi:10.1145/3737873.

[4] Y. Tian, J. K. Kummerfeld, T. J.-J. Li, T. Zhang, SQLucid: Grounding natural language database queries with interactive explanations, in: Proceedings of the ACM Symposium on User Interface Software and Technology (UIST), 2024. doi:10.1145/3654777.3676368.

[5] R. S. M. Chan, R. Sevastjanova, M. El-Assady, PleaSQLarify: Visual pragmatic repair for natural language database querying, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2026. doi:10.1145/3772318.3791265.

[6] J. Lin, R. Shan, C. Zhu, K. Du, B. Chen, S. Quan, R. Tang, Y. Yu, W. Zhang, ReLLa: Retrieval-enhanced large language models for lifelong sequential behavior comprehension in recommendation, arXiv:2308.11131, 2024.

[7] Z. Huang, Y. Zhao, W. Chen, S. Gao, K. Yu, W. Xu, M. Tang, M. Zhu, M. Xu, A natural-language-based visual query approach of uncertain human trajectories, IEEE Transactions on Visualization and Computer Graphics 26(1) (2020) 1256-1266. doi:10.1109/TVCG.2019.2934671.

[8] X. Guo, S. Yu, J. Zhang, H. Bi, X. Chen, J. Liu, A natural language-based automatic identification system trajectory query approach using large language models, ISPRS International Journal of Geo-Information 14(5) (2025) 204. doi:10.3390/ijgi14050204.

[9] H. Yanagawa, Y. Hiroi, S. Tokida, Y. Hatada, and T. Hiraki, Navigation Pixie: Implementation and empirical study toward on-demand navigation agents in commercial metaverse, IEEE Xplore, 2025. doi:10.1109/ACCESS.2025.11220445.