27th INTERNATIONAL DEPENDENCY AND STRUCTURE MODELING CONFERENCE, DSM 2025
HOBOKEN, NJ, USA, 24 – 26 September, 2025

Retrieving Asymmetrical Indirect Links Through Large Language Models

Edwin C.Y. Koh

Design and Artificial Intelligence Programme | Engineering Product Development Pillar
Singapore University of Technology and Design, Singapore

Keywords: Auto-DSM, Large language model (LLM), Retrieval-augmented generation (RAG), Dependency modelling

1 Introduction

Constructing a Design Structure Matrix (DSM) can be resource intensive and time-consuming. Koh (2024) introduced a workflow named Auto-DSM, which uses a large language model (LLM) to automatically generate a DSM based on retrieval-augmented generation (RAG). The workflow retrieves relevant data from predefined documents to determine system entities (DSM row and column headers) and their interdependencies (DSM cell entries). A prototype was developed and evaluated using diesel engine text, successfully reproducing 357 out of 462 symmetrical direct links (77.3%) when compared to a reference DSM. While the work by Koh (2024) shows promise, it is unclear whether Auto-DSM can be used to retrieve asymmetrical and indirect links.

This paper reports on an evaluation study that examines the performance of the Auto-DSM workflow in retrieving asymmetrical direct and indirect dependencies from text data comprising of plain dependency descriptions. The findings of this work can be used to support the development of automated DSM generation, enabling more advanced DSM techniques to be built on.

2 Method

This section outlines the conditions of each experimental setting and the evaluation metrics.

2.1 LLM models used

The Auto-DSM workflow uses LLM to support the automation of DSM generation. In this work, three large language models, namely GPT-4o, GPT-4, and Llama 3, were used to assess how model choice can influence Auto-DSM performance. GPT-4o (OpenAI, 2024) represents a state-of-the-art proprietary LLM that is referred to as one of OpenAI's "versatile, high-intelligence flagship models" (https://platform.openai.com/docs/models, Accessed: 14 July 2025). The specific version used in this work is "gpt-4o-2024-11-20". In contrast, GPT-4 (OpenAI, 2023) represents an older model described by OpenAI as "An older high-intelligence GPT model". The specific version used in this work is "gpt-4-0613". Llama 3 is an open-source LLM developed by Meta (Llama Team, 2024). The specific version used in this work is Ollama "llama3.3:70b".

2.2 Data used

Plain dependency descriptions of asymmetrical direct and indirect dependencies in the form of text data were used in this work so that unambiguous correct DSM answers can be manually established to check against those generated by Auto-DSM. Figure 1 shows the text data used which was taken directly from (Koh, 2022) and describes a water supply network shown in the center. The corresponding DSM is shown on the right with '1' entries representing direct links and '0' entries representing no link. Indirect links are indicated by the number of steps required for the source entity to reach the sink entity. For instance, Entity A is linked to Entity D via Entity C. Hence, the corresponding DSM cell (i.e. Column A, Row D) has a '2' entry indicating 2 steps from source to sink. Note that in most system networks, there are usually more than one path between system entities. In this case, there are less direct paths between Entity A and Entity D. For instance, Entity A is also linked to Entity D via Entity B and later Entity C (i.e. 3 steps). To avoid ambiguity, this work focuses on the most direct path between system entities.

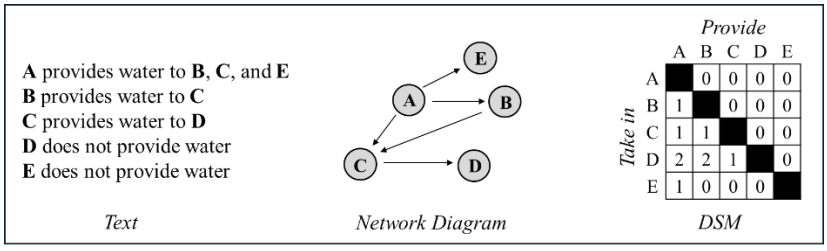

Figure 1. Text data used in this work and corresponding correct DSM answer based on (Koh, 2022).

As there are five entities A, B, C, D, and E in Figure 1, the text was reconstructed 120 times (i.e. 5 factorial = 120 permutations) to examine if the naming sequence of the entities influences the results. Each permutation was repeated 5 times resulting in 600 experimental runs for each LLM discussed in the previous section.

2.3 Prompt used

The prompt used for each Auto-DSM cell query is modified in this work as follows to capture indirect links and identify the path for each source-sink pairing:

"If <entity i> provides water to <entity j>, state the entities that form the path between <entity i> and <entity j> as a python list. Else, state [0]. Do not output anything else. If you don't know the answer, strictly state 'I do not know' instead of making up an answer."

2.4 Evaluation metrics

The DSM generated based on each text data permutation was evaluated against the corresponding correct DSM answers manually established. Five evaluation metrics were used in this work, namely Accuracy, Precision, Recall, F1 score, and Completeness. The first four are established evaluation metrics used in assessing results generated through machine learning (Powers, 2011). The last evaluation metric is adapted from (Koh, 2024) and is DSM specific, measuring the number of useful DSM cell entries generated out of the number of DSM cells to be populated.

3 Results

Table 1 shows a summary of the mean value and standard deviation (SD) of each evaluation metric with respect to the LLMs used in this work. Auto-DSM yielded a mean accuracy of 0.937 (SD = 0.026, N = 600) when GPT-4o was used. The mean accuracy increased to 0.982 (SD = 0.025, N =600) when GPT-4 was used instead. The increase in mean accuracy between GPT-4o and GPT-4 is significant based on Welch's t-test, $t(1196.90) = 30.37$, $p < 0.001$, with an effect size measured by Cohen's d as 1.75, indicating a very large magnitude of difference. The mean accuracy increased further to 0.991 (SD = 0.026, N =600) when Auto-DSM was used with Llama 3. The increase in mean accuracy between GPT-4 and Llama 3 is also significant, $t(1195.73) = 6.08$, $p < 0.001$, with an effect size measured by Cohen's d as 0.35, indicating a small to medium magnitude of difference.

Table 1. A summary of experiment results with respect to the Large Language Models used in Auto-DSM.

Model	Accuracy	Precision	Recall	F1	Completeness
GPT-4o	0.937 ± 0.026	1.000 ± 0.000	0.819 ± 0.074	0.899 ± 0.049	1.000 ± 0.000
GPT-4	0.982 ± 0.025	0.999 ± 0.013	0.949 ± 0.071	0.972 ± 0.039	1.000 ± 0.000
Llama 3	0.991 ± 0.026	1.000 ± 0.000	0.979 ± 0.057	0.989 ± 0.032	0.713 ± 0.129

In terms of precision, Auto-DSM performed better with GPT-4o and Llama 3 where each yielded a mean of 1.000 (SD = 0.000, N = 600), implying that no false positives were generated when the models were used. In contrast, Auto-DSM with GPT-4 yielded a mean precision of 0.999 (SD = 0.013, N = 600) and the decrease is statistically significant, $t(599.00) = -2.46$, $p = 0.014$, with a small effect size (Cohen's $d = -0.14$).

Similar to accuracy, Llama 3 outperformed GPT-4 and GPT-4o in terms of recall and F1 score with a mean recall of 0.979 (SD = 0.057, N = 600) and mean F1 score of 0.989 (SD = 0.032, N = 600). GPT-4 yielded a mean recall of 0.949 (SD = 0.071, N = 600) and mean F1 score of 0.972 (SD = 0.039, N = 600), while GPT-4o yielded a mean recall of 0.819 (SD = 0.074, N = 600) and mean F1 score of 0.899 (SD = 0.049, N = 600). The decrease in recall between Llama 3 and GPT-4 is significant, $t(1145.81) = -8.27$, $p < 0.001$, with Cohen's $d = -0.48$, indicating a medium magnitude of difference. The decrease in F1 score between Llama 3 and GPT-4 is also significant, $t(1152.99) = -8.33$, $p < 0.001$, with Cohen's $d = -0.48$, indicating a medium magnitude of difference. The decrease in recall and F1 score between GPT-4 and GPT-4o are significant as well, with $t(1195.86) = -30.89$, $p < 0.001$ for mean recall and $t(1141.22) = -28.62$, $p < 0.001$ for mean F1 score. The effect size is very large with mean recall Cohen's $d = -1.78$ and mean F1 score Cohen's $d = -1.65$.

While Llama 3 produced higher accuracy, recall, and F1 score, and achieved full precision (i.e. no false positives), it is worth noting that it only yielded a mean completeness of 0.713 (SD = 0.129, N = 600), implying that only 71.3% of the DSM cells were populated with useful entries on average. In contrast, the mean completeness of Auto-DSM with GPT-4 and GPT-4o is 1, implying that 100% of the DSM cells were populated with useful entries when the models were used.

Table 2 shows a summary of non-useful and erroneous entry occurrences according to categories with respect to the LLMs used. GPT-4 and GPT-4o achieved a completeness of 1 (i.e. 100% useful DSM entries) as reported in Table 1 because both did not produce non-useful results such as entries with "I do not know" or entries that are erroneous in terms of formatting or inconsistent paths. In contrast, Llama 3 only has a completeness of 0.713 (i.e. 71.3% useful entries) as it

produced a mean of 5.663 (SD = 2.589, N = 600) "I do not know" entries for each DSM generated (i.e. 5.663 out of 20 DSM cells = 28.3% non-useful entries per DSM). In addition, there were 20 occurrences of output format errors (mean = 0.033, SD = 0.180, N= 600) and 21 occurrences of path errors (mean = 0.035, SD = 0.184, N = 600) when Llama 3 was used. Examples of Llama 3 responses with format errors are shown in Figure 2, where descriptive texts that require interpretation were generated instead of concise paths written in a python list format. Format errors can be resolved by passing such responses through a second round of queries using LLMs to trim the final answer into the required format, but the approach is not explored in this paper to report the format errors in their current form. Although these responses were treated as errors in this work, the texts suggest that Llama 3 output its responses based on the most direct path available and is aligned to the assumption used in this work.

Table 2. A summary of non-useful and erroneous results with respect to the Large Language Models used.

Model	"I do not know"	Format error	Path error	Wrong Distance
GPT-4o	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.140 ± 0.392
GPT-4	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.005 ± 0.071
Llama 3	5.663 ± 2.589	0.033 ± 0.180	0.035 ± 0.184	0.212 ± 0.467

Permutation 21 using Llama 3

[A, E, C] or [A, C] Since both paths exist and we are only asked for "the entities that form the path between A and C", either of these answers would be correct. However, since A provides water directly to C according to the given information ("A provides water to E, C, and D"), the most direct path is [A, C].

Permutation 111 using Llama 3

[E, C, B] or [E, B] Since E provides water to B directly and also through C, both paths are valid. However, based on the given information that E provides water to C, B, and D, it can be inferred that E provides water to B directly as well. But since there's an alternative path through C, we acknowledge both. Given the instruction for a single answer in the format of a python list and considering direct provision is more straightforward: [E, B]

Figure 2. Examples of responses with output formating errors (i.e. descriptive text instead of concise paths in python list format).

Path errors are responses where the source and sink entities are inconsistent with the corresponding DSM cell. For example, in Permutation 25 with Llama 3, DSM cell for Entity A (source) to Entity C (sink), the Llama 3 response generated was ['B', 'A', 'C'] which erroneously reported Entity B as the source entity instead of Entity A. Such errors are harder to rectify but the number of occurrences is low.

The last column in Table 2 indicates the number of occurrences the LLMs correctly indicated a link between entities but did not indicate the most direct path, resulting in a wrong path distance. For instance, Figure 3 shows the text processed by Auto-DSM for Permutation 53 and the intended response based on the most direct paths. The DSM generated based on Auto-DSM with GPT-4o is shown on the right (Trial 3). The intended response from Entity C to Entity B is '2' steps based on the direct path "C to E to B" inferred from the text. However, the corresponding GPT-4o response was '3' steps based on the path "C to A to E to B", resulting in a wrong distance error. Such occurrences happened with a mean of 0.140 (SD = 0.392, N = 600) for each DSM generated using GPT-4o. The mean decreased to 0.005 (SD = 0.071, N = 600) when GPT-4 was used instead, and the difference was significant, $t(637.73) = -8.29$, $p < 0.001$, with a medium effect size (Cohen's $d = -0.48$). The occurrences increased when Llama 3 was used, resulting in a mean of 0.212 (SD = 0.467, N = 600). The increase in mean occurrences of wrong distance error between GPT-4 and Llama 3 is significant with $t(626.47) = 10.74$, $p < 0.001$, with a medium to large effect size (Cohen's $d = 0.62$).

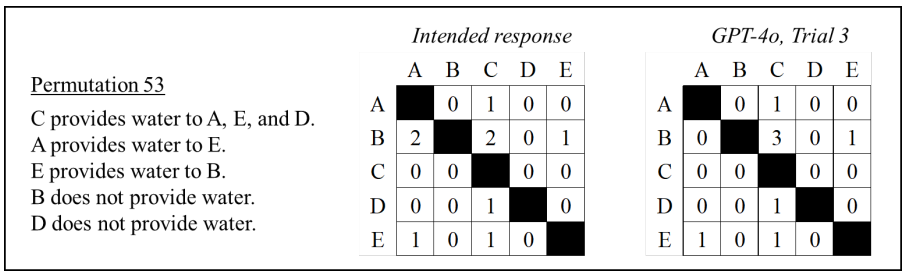

Figure 3. Text used for Permutation 53, the intended DSM, and a corresponding DSM generated using Auto-DSM with GPT-4o.

5 Closing remarks

This paper explores the use of Auto-DSM on text documents with plain dependency descriptions to examine the potential of retrieving asymmetrical indirect links for automated DSM generation. The work was repeated using three large language models, namely GPT-4o, GPT-4, and Llama 3, to assess how model choice influence performance evaluation metrics. The results show that GPT-4 outperformed or tied with GPT-4o across evaluation metrics (except for precision) even though GPT-4 is an older model compared to GPT-4o. In addition, even though Llama 3 outperformed GPT-4 in accuracy, precision, recall, and F1 score, Llama 3 only achieved a mean of 71.3% completeness in useful DSM cell entries compared to 100% completeness when GPT-4 was used, highlighting the practical tradeoffs in automated DSM generation.

In this work, Auto-DSM achieved an accuracy of 0.937 and above, suggesting that asymmetrical indirect links can be retrieved using Auto-DSM for automated DSM generation. The result is higher than the 77.3% accuracy reported in (Koh, 2024), which used text data in formats such as textbooks instead of plain dependency descriptions. The findings show promise and suggest that data format can have a strong influence on Auto-DSM performance.

References

Koh, E.C.Y., 2022. Resilience analysis of infrastructure systems in incremental design change. Computers in Industry 142, 103734.

Koh, E.C.Y., 2024. Auto-DSM: Using a large language model to generate a design structure matrix. Natural Language Processing Journal 9, 100103.

Llama Team, 2024. The llama 3 herd of models. arXiv. https://doi.org/10.48550/arXiv.2407.21783

OpenAI, 2023. GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

OpenAI, 2024. GPT-4o system card. arXiv. https://doi.org/10.48550/arXiv.2410.21276

Powers, D.M.W., 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning Technologies 2 (1), pp 37-63.

Contact: Edwin C.Y. Koh, DAI/EPD, Singapore University of Technology and Design, edwin_koh@sutd.edu.sg