Supplementary Material S1: Analyses on Prior LLMs

S1. Quantitative Evaluation

S1.1. Overview

In this Supplementary Material, we examine the scores produced by well-known LLMs that used to be state-of-the-art. We selected GPT-4o and Claude 3.5 Sonnet (20240620) as they were previously the leading solutions for multimodal learning (`https://arena.ai/leaderboard/vision`, rank on July 1st 2024). We also included deepseek R1 7B distilled combined with Janus Pro 1B, as the DeepSeek suite of AI tools received significant attention following its release in 2025.

As in the main manuscript, the scores are performed for all three ABMs and we vary the content of the prompts using a factorial design of experiments such that we can report on interactions. As stated in the main manuscript, our quantitative evaluation reports six scores: BLEU [1, 2], METEOR, ROUGE-1, ROUGE-2, ROUGE-L [1, 3], and the Flesch Readability Score [4]. While the last score does not need a reference, the first five scores are computed by comparing the LLM's output with a reference summary. This complementary assessment makes four simplifications compared to the main manuscript. Since the evaluation focuses on the ability of LLMs to detect important trends in the visuals, we provide neither statistical summaries nor the output produced by the context stage to help the LLMs. The evaluation accounts for both extractive (BERT) and abstractive (BART) summarization algorithms and we evaluate results with respect to one summary for each ABM (labeled as 'author' in **Supplementary Material S2**).

S1.1.1. Optimal performances

The best results reported in Table 1 show that performances heavily depend on the experimental setting and on the preferred metric, as different LLMs optimize different aspects of the text and react differently across ABMs. Despite this volatility, we see that Claude is often the best choice (4 out of 6 times) for reading ease while GPT is often the best choice (4 out of 6 times) for most of the BLEU, ROUGE, and METEOR metrics. The Flesch readability ease rewards shorter words, shorter sentences, and simpler syntax, which are aligned with Claude's tendency to generate less formal structures in conversational tone. In contrast, GPT is heavily optimized for downstream NLP benchmarks such as captured by BLEU, ROUGE, and METEOR, where being conservative in paraphrasing is an asset. DeepSeek is a surprisingly competitive solution, particularly when considering that it is a distilled model with 'only' 7 billion parameters, as it does very well for the RAGE ABM (sometimes outperforming the other two LLMs by a wide margin) but is weaker on the other two ABMs. The intermediate positioning of DeepSeek was

Email address: `*pgiabban@odu.edu` ()

confirmed by our t-tests, as there is a statistically significant difference between DeepSeek and GPT on reading ease (p=0.0035, t=3.00) and also a statistically significant difference between DeepSeek and Claude on METEOR (p=0.0020, t=3.18).

When considering the summarization algorithms, BART produces summaries closer in phrasing and structure to the human-written summaries. This suggests that summaries written by humans are not copied from source documents but rather rewritten or paraphrased by modelers to provide a high-level picture. To contextualize the influence of the ABMs, note that MHPM is about ecological and evolutionary modeling, which may be more technical and abstract. This technical language and complexity make the summaries harder to simplify and they may favor LLMs that can reproduce domain-specific vocabulary. The Milk Adoption ABM is about behavioral modeling with a consumer focus, discussing decision-making, preferences, and historical trends. This more familiar content lets GPT match wording and Claude produce smooth, easy-to-read text. DeepSeek's proficiency in handling the RAGE model potentially arises from its capacity to process and summarize content characterized by high variability, agent heterogeneity, and adaptive behaviors.

Table 1: For each ABM, LLM, and summarization algorithms, we report the best result based on optimizing the prompt. For brevity, ROUGE is denoted by R.

Experimental condition			**Best score across dynamic prompt elements**					
ABM	***Summary***	***LLM***	BLEU	METEOR	R-1	R-2	R-L	Reading ease
Milk	BART (abstract)	Claude	0.11	0.35	0.58	0.22	0.27	63.70
		GPT	0.13	0.36	0.64	0.27	0.28	54.83
		DeepSeek	0.09	0.34	0.59	0.22	0.25	61.87
	BERT (extract)	Claude	0.09	0.28	0.57	0.21	0.24	59.84
		GPT	0.10	0.29	0.59	0.22	0.24	59.03
		DeepSeek	0.06	0.26	0.56	0.20	0.24	59.33
RAGE	BART (abstract)	Claude	0.03	0.23	0.42	0.10	0.17	57.27
		GPT	0.03	0.26	0.44	0.11	0.18	55.44
		DeepSeek	0.04	0.30	0.50	0.17	0.20	56.25
	BERT (extract)	Claude	0.03	0.19	0.40	0.09	0.16	52.09
		GPT	0.02	0.20	0.42	0.09	0.17	51.99
		DeepSeek	0.04	0.25	0.47	0.12	0.21	60.04
MHPM	BART (abstract)	Claude	0.03	0.25	0.40	0.10	0.17	54.52
		GPT	0.04	0.25	0.41	0.10	0.17	47.69
		DeepSeek	0.02	0.27	0.33	0.07	0.15	45.46
	BERT (extract)	Claude	0.04	0.21	0.40	0.10	0.17	51.07
		GPT	0.04	0.21	0.42	0.10	0.17	51.18
		DeepSeek	0.03	0.25	0.40	0.08	0.17	51.18

S1.1.2. Impact of Experimental Design Factors

Table 2 reports statistically significant effects of an experimental variable (including the three dynamic prompt elements) across the metrics. We observe that the content and structure of the underlying ABM has a major influence on text generation quality – likely due to differences in domain complexity, vocabulary, and narrative style. The summarization technique has a clear impact on text fidelity and linguistic quality, especially for deeper n-gram overlap and ease of

reading. The choice of LLM has a consistent effect on output, especially regarding precision and clarity. In contrast, the dynamic prompt elements show little to no statistical significance across metrics *except for the reading ease*.

Table 2: p-value from a one-way ANOVA between our six experimental variables (including the three dynamic prompt elements in the bottom rows) and our six quantitative metrics, where 'R' denotes ROUGE.

↓*Variable / metric* →	BLEU	METEOR	R-1	R-2	R-L	Reading ease
Agent-Based Model	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
Summarization algorithm	<0.01	<0.01	0.09	<0.01	<0.01	0.71
LLM	<0.01	<0.01	<0.01	0.10	0.21	<0.01
Use of roles	0.83	0.89	0.96	0.35	0.42	0.36
Generating insights	0.45	0.30	0.37	0.73	0.65	<0.01
Providing examples	0.76	0.49	0.53	0.32	0.26	<0.01

Unlike ANOVA, which only tests for statistical significance, the factorial analysis summarized in Table 6 quantifies the relative contribution of each factor to total variance, also making it possible to compare their *joint influence* on model performance. That is, we show *dependencies in the architecture* as the same prompt yields different outcomes depending on the LLM or summarizer used. While these results confirm the role of the summarization algorithm and the LLM, they provide additional information to the ANOVA. The summarization algorithm (BART vs. BERT) determines how the ABM outputs are reworded and what information is emphasized, which impacts metrics that directly measure lexical overlap (BLEU, METEOR, and ROUGE scores). So the choice of summarization algorithm clearly dominates the variance in these metrics, especially for METEOR, which also accounts for paraphrasing and synonyms (where BART's generative style may shine more). As we know from the ANOVA, the LLM choice was statistically significant for BLEU and ROUGE-1, which are sensitive to surface-level lexical overlaps where LLMs clearly differ (e.g., Claude paraphrasing more, GPT being more literal). The LLM does not seem to affect METEOR because the summarization algorithm dominates. *Most interestingly, the factorial analysis reveals that prompt elements do matter either individually or by interacting with the other technical choices*. Asking for insights has a noticeable impact on reading ease since it affects how the LLM frames the outputs by using explanatory language and potentially staying away from dense or technical phrasing. Interactions between insights and examples emerge because examples help ground the response while insights elevate reasoning, leading to a small improvement on n-gram matches (ROUGE-1) with reference summaries written in similar style. The interactions of LLMs and examples for ROUGE-1 and ROUGE-2 suggest that LLMs interpret example-driven prompts in different ways (e.g., Claude may elaborate more). The interactions of summarization algorithms and insights for BLEU potentially reflect that requests for insights generate more reasoning or causal language, which can be captured more fluidly in abstractive summaries than in extractive summaries.

The correlation matrix (Table 4) confirms expected relationships among evaluation metrics: BLEU, METEOR, and ROUGE variants are strongly intercorrelated due to their shared reliance on lexical overlap, while Flesch reading ease shows weaker, moderate correlations, reflecting its distinct focus on stylistic clarity rather than content similarity. These correlations also support our results in section 4.2.2. On the one hand, we observe one cluster of content metrics (BLEU, METEOR, ROUGE) that are strongly correlated, thus validating their shared reaction to the summarization algorithms. On the other hand, we see the readability measure, which is weakly

Table 3: Effect of individual prompt elements and pairs of elements onto the quantitative scores. This uses a factorial design of experiments, thus the scores correspond to the percentage of contribution to variance. Scores above 5 are emphasized **in bold.**

Feature	BLEU	METEOR	R-1	R-2	R-L	Reading ease
Example	1.85	0.39	**5.09**	0.56	0.90	**19.32**
Example *and* Insight	0.07	0.56	**5.49**	2.05	2.02	0.10
Insight	0.30	0.17	2.52	4.63	**5.66**	**43.64**
LLM	**29.07**	2.70	**32.87**	3.06	0.26	**21.57**
LLM *and* Example	0.24	0.47	**9.98**	**6.94**	1.10	4.68
LLM *and* Insight	0.19	0.08	0.58	0.28	0.74	3.15
LLM *and* Role	1.95	0.08	0.13	0.81	0.69	1.94
Role	0.14	0.01	0.01	4.01	2.91	1.47
Role *and* Example	0.43	0.01	0.42	3.21	0.49	1.21
Role *and* Insight	1.15	0.11	1.21	0.46	0.07	1.86
Summary	**51.29**	**94.93**	**18.72**	**72.44**	**55.97**	0.25
Summary *and* Example	0.62	0.33	0.52	0.30	1.63	0.03
Summary *and* Insight	**6.93**	0.12	0.00	0.30	0.31	0.63
Summary *and* LLM	**5.27**	0.04	**20.97**	0.21	**26.32**	0.02
Summary *and* Role	0.50	0.02	1.50	0.75	0.93	0.14

correlated to the rest, thus validating its separate reaction to prompt design and the LLM.

Table 4: Correlation matrix among the quantitative scores to verify expected relationships.

	BLEU	**METEOR**	**R-1**	**R-2**	**R-L**	**Reading ease**
BLEU		0.76	0.91	0.94	0.93	0.47
METEOR	0.76		0.70	0.76	0.77	0.35
R-1	0.91	0.70		0.96	0.96	0.55
R-2	0.94	0.76	0.96		0.9	0.55
R-L	0.93	0.77	0.96	0.96		0.54
Reading ease	0.47	0.35	0.55	0.55	0.54	

S2. Qualitative Evaluation

S2.1. Principles of Qualitative Evaluation

Qualitative evaluation involves assessing the generated content on criteria such as *faithfulness* and *coverage*. While *fluency* (quality of writing) was measured in earlier studies as GPT-2 and GPT-3 occasionally struggled with grammar or spelling [5], this is less of a concern with the state-of-the-art LLMs available in 2025. Faithfulness refers to how accurately a summary reflects the facts and meaning of the source content without introducing errors or hallucinations, while coverage measures how well the summary includes the key points and important information from the original text. In other words, we would like key data from the simulation to be in the summary (coverage) and we would not like content in the summary that wasn't sourced from the simulation (faithfulness).

The evaluation can be *performed manually by human readers*. A structured process ensures consistency and reliability [6]: initially, a first batch of summaries are rated collaboratively to establish a standard and negotiate meaning. Subsequently, another batch is independently rated to measure interrater agreement using Cohen's Kappa, which ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond chance, and negative values indicate agreement less than chance[1]. Finally, if a satisfactory level of interrater agreement has been achieved, the remaining results can be individually rated, thus allowing to cover a larger amount of data for the same manpower.

While qualitative assessment frequently relies on human evaluators, it can *also be conducted with LLMs*. This allows the analysis to be performed at a larger scale, which may be needed when scanning a large volume of summaries generated by LLMs across various parameter values. When LLMs are used, Cohen's Kappa is still computed between the LLM and a human reader to assess whether the LLM is capable of independently rating the material.

S2.2. Calibration process and Kappa scores

We explored both qualitative rating by human readers and evaluation by LLMs.

First, we hired two human readers, external to the authors of this manuscript to avoid conflicts of interest by which we would develop and rate our own solutions. Using the process outlined in the previous section, the two human readers perform a qualitative evaluation by ranking the generated reports according to faithfulness and coverage using the 5-point Likert scale [8, 9]. Since LLM generations often involve a large amount of output, it is common to manually assess only a random sample of these results. We provided a rubric and walked the raters on how to rate summarized reports for faithfulness and coverage based on a simulation model that was not used in our study. Then, we asked the raters to independently assess 6 (out of 432) of our generated reports. We computed their interrater agreement using the Cohen's Kappa metric

Since the 5-point Likert scale consists of ordered categorical variables comprising multiple categories, we used the weighted Kappa as recommended in the literature [10]. The results from this first batch were $\kappa_{coverage} = 1$ and $\kappa_{faithfulness} = 0.33$. We facilitated a meeting with the two raters, asking them to justify their assessments and come to an agreement on ratings for all cases that were interpreted differently. The raters often disagreed on the scoring and only occasionally reached consensus after discussion. After rating additional batches of reports and meeting again two more times, the kappa scores *declined*. The *decline in Cohen's kappa following the adjudication process likely reflects the inherent complexity* and subjectivity of the evaluation task. Indeed, it is expected that tasks requiring inference (e.g., generating insights from simulation outputs and connecting them to a model's design and goals) lead to significantly lower agreement compared to simple tasks [11]. As pointed out in the main manuscript, the fact that different human readers do not agree on what should go in a summary of a simulation model is in line with prior works, in which several human readers were given the same summaries yet converted them into different models [12].

[1] A common challenge is the high variability in Cohen's Kappa values, indicating inconsistent agreement among raters: some tudie report a very high kappa of 0.93 [6], while other may have much lower value. To address this, researchers are increasingly defining detailed rubrics to enhance precision and clarity in ratings. Two predominant approaches are often employed: descriptive rubrics [7], which provide detailed criteria for each level of performance, and holistic rubrics [6], which offer a single overall score based on a general impression of performance. These rubrics aim to standardize the evaluation process, reduce variability, and ensure more reliable and consistent assessments across different raters

Second, we evaluated the ability of three LLMs (DeepSeek's Janus Pro 1B, GPT-4o, Claude 3.5 Sonnet) at evaluating faithfulness and coverage. We computed the inter-rater agreement between these solutions and a human rater. The scores shown in Table 5 suggest two takeaways: (i) no single LLM is best at evaluating all aspects, and (ii) the extreme scores (e.g., GPT thinks faithfulness is always poor while Claude believes that coverage is usually perfect) and low inter-rater agreement limit the interpretability of evaluations by LLMs. The outputs from Janus Pro 1B were not usable as it hallucinated to the point that it did not provide scores.

Table 5: Ability of different LLMs at evaluating faithfulness and coverage with respect to a human rater.

	Faithfulness			**Coverage**		
Cases	Human	GPT	Claude	Human	GPT	Claude
1	1	2	5	5	4	5
2	2	2	5	5	3	5
3	1	1	2	4	1	2
4	3	2	5	5	4	5
5	5	2	5	5	3	5
6	3	2	4	5	3	5
7	4	3	5	5	4	5
8	4	2	5	4	2	5
Weighted kappa		0.2	0.16		0	0.33
% agreement (Po)		0.71	0.59		0.56	0.90

S2.3. Results

While noting the limitations stated in the subsection above, we have provided the qualitative scores for a comprehensive assessment-by-LLM of all automatically generated reports. Note that *using an LLM to evaluate its own output is risky*, especially for tasks requiring factual correctness or critical judgment. Instead of self-evaluating, the LLM tends to favor its own responses, lacks objectivity, and may confuse fluency with accuracy. For example, it would be problematic to directly generate a summary via GPT and ask GPT whether it likes its own summary. In our case, the risk is mitigated because the text *initially* generated by a LLM is *summarized* through another algorithm (BERT or BART) and then it is evaluated.

The inter-rater agreements showed us that GPT is best for faithfulness while Claude is preferable for coverage. However, GPT tends to give low scores while Claude yields high scores. The evaluation by LLMs is still informative because, even if the absolute score is overly generous or severe, it responds to changes in our experimental design (i.e., the prompt elements such as roles and insights). In other words, we do not use the LLMs to make absolute judgments ("this summary is bad") but rather to detect relative differences ("we have higher scores if the prompt asks to generate insights"). The factorial analysis in Table 6 shows that the choice of LLM makes the most difference. Interestingly for coverage, some LLMs are better able to utilize examples than others and setting a role has an impact. When it comes to faithfulness, we observe interactions between setting a role and providing examples or asking for insight generation.

S3. Examples of complete prompts

Prompt 1 : `text` is a feature from the design of experiments. *text* is a part of the prompt changing according to the model and the experiment

`You are an expert in Consumer Behavior without any statistics background.`

Your goal is to explain an agent-based model. Your explanation must include the context of the model, its goals, and key parameters. For each parameter, include the range of values (if provided) and the value that we have set. Do not write any summary or conclusion.

We have 16 parameters:

- *memory-lifetime, from* 1.0 *to* 10.0*. We set it to* 5*.*
- *alt-health-mean-initial, from* 1.0 *to* 3.0*. We set it to* 1.69*.*
- *habit-threshold, from* 1.0 *to* 10.0*. We set it to* 2*.*
- *p-interact, from* 0.0 *to* 1.0*. We set it to* 0.3*.*
- *social-susceptibility, from* 0.0 *to* 1.0*. We set it to* 0.2*.*
- *incumbent-initial-habit, from* 0.0 *to* 10.0*. We set it to* 10*.*
- *alt-env-mean-initial, from* 0.0 *to* 3.0*. We set it to* 1.12*.*
- *social-blindness, from* 0.0 *to* 1.0*. We set it to* 0.51*.*
- *justification, from* 0.0 *to* 1.0*. We set it to* 0.61*.*
- *cognitive-dissonance-threshold, from* 0.0 *to* 1.0*. We set it to* 0.15*.*
- *network-parameter, from* 2.0 *to* 10.0*. We set it to* 8*.*
- *number-of-agents. We set it to* 1000*.*
- *habit-on?. We set it to True.*
- *network-type. We set it to "watts-strogatz".*
- *norms. We set it to True.*
- *social-conformity. We set it to* 0.28*.*

The context and goals of the model are as follows.
The overall objective of the agent-based model is to reproduce adoption behaviours of milk consumption by the British public (1974-2005), by replicating individual preferences and decision factors. Specifically, the goal of the model is to explore the influence of perception, habits and social influence in an individual's decision-making process of milk choice. The model looks to replicate the substitution of whole milk for skimmed and semi-skimmed milk from the 1970s onwards. The main outcome reported here is the average weekly consumption of each milk type per person. The simulation uses both a theoretical grounding and empirical data to inform the ABM, with calibration performed against observed macro level data.
In particular, the study conducts an experiment to compare the performance of two model variants in reproducing observed milk consumption trends. These variants present different mechanisms for how agents become disposed to consider their choices, representing a threshold based, and a probability-based approach. Decision-making follows a basic structure of: agent perception of choice characteristics, the triggering, or not, of disposition to consider alternatives, a set of scored choice functions made up of the perceived characteristics and modulated by habit and social influence, and finally, choice

evaluation where agents consider the impact of their choices and may adjust their future decisions.
The agents in the model represent adult consumers who occupy a random position in an information environment. Each agent has a disposition to consider alternative milk choices. Two disposition mechanisms are tested in the model, a threshold-based approach, and a probability-based approach. Each agent makes a choice of milk selection based on a function for each alternative, made up of the perceived health and environmental characteristics of each choice. These are computed at the initialisation of the simulation and then calculated at each time step. An agent's milk choice function is modified by habit, social influence, and evaluation of previous choices. Agents ascribe different relative importance to each constituent part of the choice function (health factors and environmental factors). If the disposition requirement has been met, consumption of each milk type is split proportionately by the size of each choice function, modulated by the other influences. If not, agents keep their existing choice.
Agents (n=1,000) start with an existing choice based on the dominant position of whole milk versus skimmed varieties in 1974 (start year of the data). All agents are part of a social network. Each agent in the network can sense and be influenced by the choice function of each milk alternative for other agents in their network. Links between agents are unidirectional, and influence occurs as a function of interaction probability, with the degree of influence characterised by agent susceptibility. Social norms are globally perceived by agents and impact the weightings of the choice function Setup and Go.
Monitors:
Two monitors track the number of turtles that have consumed mostly whole-milk (incumbent initial choice) or skimmed/semi-skimmed milk (alternative initial choice).
Plots:
One plot shows the percentage of majority whole milk or skimmed/semi-skimmed milk consumers.
The second plots the average consumption of each milk type over time. This is the data that the calibration exercise uses to try and replicate observed consumption data.
Parameters:
12 parameters that governs whether an agent will consider changing its milk type, and how much of each milk type an agent will consume. We refer viewers to the manuscript based on this model for further details. The appearance of a crossover in the types of milk consumed and the shape of the curves over time.

Prompt 2 : `text` is a feature from the design of experiments. *text* is a part of the prompt changing according to the model and the experiment

```
You are an expert in Consumer Behavior with a statistics background
```

We have a plot from repeated simulations of an agent based model. Your goal is to describe the trends in details from the plot, mentioning key time steps and values taken by the metric in the plot, and interpreting them based on the context of the model. The report must objectively describe the trends in the data without addressing the quality of the simulation. Do not refer to the plot or any visual in your description. If a plot has very simple dynamics, simply state them without expanding.

```
Here is an example of the style of the report: The total number
of earthquakes recorded in the region starts at 5,000 then it
declines rapidly over the first 400 time steps. This initial
decline could be attributed to the depletion of immediate stress
points within the fault lines, as the most susceptible areas
release their pent-up energy early in the simulation. It increases
briefly at around 500 steps, likely due to the aftershocks and
secondary stress points being triggered as the system seeks a new
equilibrium. The simulation ends with the number of earthquakes
near zero, indicating that the majority of the stress has been
released and the system has stabilized.

The total seismic activity follows the same pattern, starting
at 10,000 events and declining steadily. This steady decline
reflects a systematic reduction in seismic energy over time, as
the energy distribution within the tectonic plates becomes more
uniform. There is low variance across simulation runs in the
first 100 steps, but deviations become noticeable afterward. This
suggests that while the initial reactions to stress are highly
predictable, as time progresses, the system's complexity introduces
more variability. This variability could be due to differences in
secondary stress points and the non-linear nature of seismic energy
release over time.

When summarizing trends, provide brief insights about their
implications for decision makers.
```

The attachment is the plot representing the average weekly consumption of whole milk per agent

References

[1] X. Zhang, Y. Li, J. Wang, B. Sun, W. Ma, P. Sun, M. Zhang, Large language models as evaluators for recommendation explanations, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 33–42.

[2] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[3] L. Chin-Yew, Rouge: A package for automatic evaluation of summaries, in: Proceedings of the Workshop on Text Summarization Branches Out, 2004, 2004, pp. 74–81.

Table 6: Effect of individual prompt elements and pairs of elements onto coverage and faithfulness. This uses a factorial design of experiments, thus the scores correspond to the percentage of contribution to variance. Pairs that are not displayed (e.g., case study and summary) were omitted for brevity since both scores were 0.

Feature	**Coverage (Claude)**	**Faithfulness (GPT)**
Case study	5.10	2.34
Case study *and* Example	0.30	1.77
Case study *and* Insight	0.44	0.99
Case study *and* LLM	4.32	1.62
Case study *and* Summary	1.59	4.06
Example	0.01	0.17
Insight	0.29	0.02
Insight *and* Example	0.15	0.08
LLM	33.57	32.06
LLM *and* Example	11.80	1.85
LLM *and* Insight	0.05	0.50
LLM *and* Summary	0.62	1.89
Role	3.16	0.02
Role *and* Case study	0.51	1.63
Role *and* Example	0.01	0.00
Role *and* Insight	0.15	1.93
Role *and* LLM	1.34	0.73
Role *and* Summary	0.29	0.95
Summary	0.01	0.31
Summary *and* Example	0.72	0.02
Summary *and* Insight	0.05	0.48

[4] S. R. Cooperman, R. A. Brandão, Investigating the proficiency of an ai tool in summarizing foot and ankle literature: A quantitative, qualitative and accuracy analysis, Foot & Ankle Surgery: Techniques, Reports & Cases 4 (2) (2024) 100384.

[5] A. Shrestha, K. Mielke, T. A. Nguyen, P. J. Giabbanelli, Automatically explaining a model: Using deep neural networks to generate text from causal maps, in: 2022 Winter simulation conference (WSC), IEEE, 2022, pp. 2629–2640.

[6] B. Naismith, P. Mulcaire, J. Burstein, Automated evaluation of written discourse coherence using gpt-4, in: Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), 2023, pp. 394–403.

[7] S. Moore, H. A. Nguyen, T. Chen, J. Stamper, Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods, in: European conference on technology enhanced learning, Springer, 2023, pp. 229–245.

[8] T. Fischer, S. Remus, C. Biemann, Measuring faithfulness of abstractive summaries, in: Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), 2022, pp. 63–73.

[9] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170.

[10] M. Li, Q. Gao, T. Yu, Kappa statistic considerations in evaluating inter-rater reliability between two raters: which, when and context matters, BMC cancer 23 (1) (2023) 799.

[11] P. Raghavan, E. Fosler-Lussier, A. M. Lai, Inter-annotator reliability of medical events, coreferences and temporal relations in clinical narratives by annotators with varying levels of clinical expertise, in: AMIA Annual Symposium Proceedings, Vol. 2012, 2012, p. 1366.

[12] A. J. Freund, P. J. Giabbanelli, Are we modeling the evidence or our own biases? a comparison of conceptual models created from reports, in: 2021 annual modeling and simulation conference (ANNSIM), IEEE, 2021, pp. 1–12.