Introduction:
Models for assessing aphasia have moved from an impairment-based to a participation-based framework [1], emphasizing tools to measure improvements in real-world communication. Several tools within this paradigm [2-4] evaluate transactional success in conversation by having persons with aphasia (PWA) retell stories co-constructed with non-aphasic conversation partners (CPs). Subsequent story retells by CPs are then analyzed by measuring how many main concepts (MCs) —propositions that capture the story's gist—are successfully conveyed between the PWA and CP. Currently, MCs are predetermined through a labor-intensive human curation process [2-4], limiting a tool's widespread clinical adoption and adaptation to new stimuli. This study aims to automate the generation of MCs for new, personalized stimuli, to make these assessment tools more feasible for clinicians and clinical researchers and more relatable for patients.

Aim: The surge in the application of large language models (LLMs) presents a novel opportunity to streamline discourse analysis in aphasia testing and outcome measurement. We aim to investigate LLMs' potential to automatically generate candidate main concepts (MCs) for novel personalized stimuli that can be reviewed by clinical researchers. In this study, we first examine whether LLMs can generate all the essential main concepts relevant to a stimuli, by comparing LLM-generated MCs with human-curated "gold-standard" MCs. We further assess whether the list of generated MCs is concise for practical manual review.

Methods:
Dataset: For gold-standard MCs, we use existing dataset [3], which contains short video/audio clips as stimuli, each annotated with MCs. The stimuli were designed to evoke emotional responses and produce substantial story retells. MCs were obtained from non-aphasic participants' retellings of stimuli, with each stimulus having 7-12 MCs (Table 1). Each MC is a statement with one main verb, and its subject, object, modifiers, and subordinate clauses if appropriate [5].

Two-stage MC generation method:
Stage 1: To generate MCs, we prompt-engineered *Llama-3-70B*, one of the best-performing open-sourced LLMs [6]. Since LLMs outputs vary based on prompt phrasing [7], we experiment with different prompts. Prior work often selects the best-performing prompt, but we found this insufficient as different prompts can capture different story aspects. To address this, we combine LLM outputs from three best-performing prompts for better MC coverage (Table 2). To ease manual review of the generated MCs, we further prompt the LLM to break down its outputs into statements that match gold-standard MC structure. Finally, to assess the quality of LLM-generated MCs, we prompt-engineered GPT-4 to compare them with gold-standard MCs and compute recall—fraction of gold-standard MCs present in the LLM-generated MCs.

Stage 2: While a naive union of outputs from different prompts can improve recall, it can create a long list with many redundant MCs that are similar in meaning but expressed using different words. This results in a very high yield (total words in MC list), making manual review cumbersome. We address this by grouping MCs that are similar in meaning and selecting one representative MC from each group. We first convert each MC into a numerical vector using Sentence-BERT [8], designed to represent sentences with similar meanings as similar vectors. We then cluster these vectors using DP-means clustering algorithm [9]. In

this algorithm, a user-specified distance threshold (δ) controls how tight or spread apart clusters are—since vectors closer to each other tend to have similar meaning, a low δ groups highly similar or near-duplicate MCs, while a high δ also groups moderately similar MCs. From each cluster, we select the MC vector closest to the center and use its text as the representative MC. The representative MCs form our final MC list, which we evaluate using recall and yield.

Results:
Figure 1 shows that union over outputs from different prompts (stage 1) obtains a high recall (0.92), suggesting that most of the gold-standard MCs are generated by the LLM. In contrast, individual prompts achieve much lower recall (0.66-0.69), demonstrating the benefit of combining outputs from different prompts. Clustering (stage 2) helps reduce yield with only a small drop in recall. For instance, at δ=0.4, clustering results in a yield of 634 words, much smaller than the naive union (yield=2933.6 words), while maintaining a high recall (0.87).

Discussion:
Our results demonstrate LLMs' strong potential in generating reliable, concise MCs comparable to human raters. We are actively investigating these strong early results, aiming to replicate our findings using different stimuli and alternate LLMs. Our approach can allow adaptation of many existing assessment tools to novel stimuli personalized for patients with different cultural backgrounds, potentially revolutionizing the aphasia intervention landscape.

References

[1] Brady, M. C., Kelly, H., Godwin, J., Enderby, P., & Campbell, P. (2016). Speech and language therapy for aphasia following stroke. Cochrane Database of Systematic Reviews, 2016(6). Article No. CD000425.

[2] Carragher, M., Sage, K., & Conroy, P. (2015). Preliminary analysis from a novel treatment targeting the exchange of new information within storytelling for people with nonfluent aphasia and their partners. Aphasiology, 29(11), 1383–1408. https://doi.org/10.1080/02687038.2014.988110

[3] Kurland J, Liu A, & Stokes P (2021). Phase I test development for a brief assessment of transactional success in aphasia: Methods and preliminary findings of main concepts in non-aphasic participants. Aphasiology, 37, 39-68.

[4] Ramsberger, G., & Rende, B. (2002). Measuring transactional success in the conversation of people with aphasia. Aphasiology, 16(3), 337–353. https://doi.org/10.1080/02687040143000636

[5] Richardson JD, & Dalton SG (2016). Main concepts for three different discourse tasks in a large non-clinical sample. Aphasiology, 30(1), 45-73.

[6] Llama team. The Llama 3 Herd of Models (2024). ArXiv, abs/2407.21783. https://llama.meta.com/

[7] Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A (2024). Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. Twelfth International Conference on Learning Representations.

[8] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Conference on Empirical Methods in Natural Language Processing.

[9] Dinari, O. & Freifeld, O. (2022). Revisiting DP-Means: Fast scalable algorithms via parallelism and delayed cluster creation. Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence, PMLR 180:579-588, 2022. https://proceedings.mlr.press/v180/dinari22b.html.

[10] Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. Cognitive Psychology, 9, 111–151.

[11] Richardson, J. (2009). The next step in guided reading: Focused assessments and targeted lessons for helping every student become a better reader. Scholastic Incorporated.

Figures and Tables

#Title	Condition	time (s)	# MCs	Description	Source
Marcus Yam	VS	198	11	photojournalist	PBS "Brief but Spectacular" series
Sylvia Earle	VS	181	8	marine biologist	PBS "Brief but Spectacular" series
Naomi DeLaRosa	VS	194	8	on family separation	PBS "Brief but Spectacular" series
Robin Steinberg	VS	128	7	the bail project	PBS "Brief but Spectacular" series
Ferguson	SD	178	10	Ferguson protesters find friendship	NPR "StoryCorps" series
Sep 11	SD	172	12	one survivor's story	NPR "StoryCorps" series
Aunt Mother	SD	166	7	aunt turned mother after tragedy	NPR "StoryCorps" series
No Handbook	SD	156	11	mother/son discuss school shootings	NPR "StoryCorps" series

Table 1: Descriptive information on stimuli used in this work. VS=visually supported biographical video; SD=speech-dependent audio clip with only a single still photo for visual support.

Prompt	Recall
Please provide a list of main concepts mentioned in the story using the five finger retell strategy.	0.63
Please provide a list of main concepts mentioned in the story using the 7-unit story grammar.	0.59
Please provide a list of main concepts mentioned in the story using the 5 W's, 1 H retell strategy.	0.47
Please provide a list of main concepts mentioned in the story using the 5 W's retell strategy.	0.50
Suppose you are a story reteller, generate a list of main concepts to help me understand the story.	0.56
Suppose you are a story reteller. Please identify one plot point in the given story in 30 words or less.	0.28
Suppose you are a story reteller. Please identify two plot points in the given story in 60 words or less.	0.39

Table 2: List of different prompts we experimented with and the corresponding recall (fraction of gold-standard MCs present in the LLM-generated MCs). We consider prompts that mention names of different theories used for understanding story structure [10] or strategies used to teach story retelling to students [11]. In each of these prompts, we also provide a definition of the theory (omitted here for brevity). We also experiment with other simpler prompts asking the LLM to help understand the story or explain the plot points of the story.

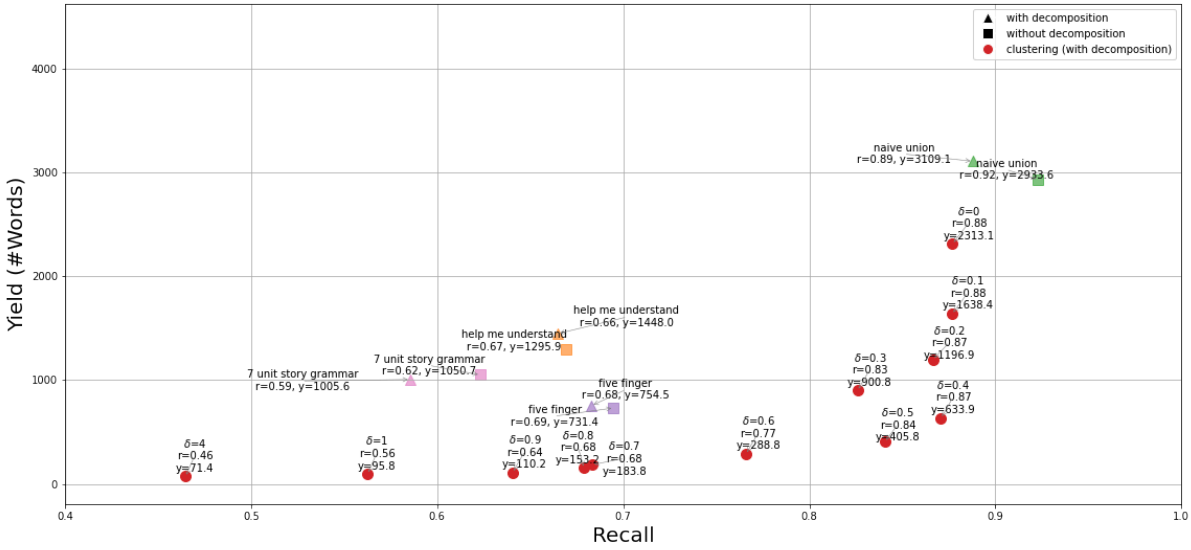

Figure 1: The recall (r) versus yield (y) tradeoff for different prompting methods used for generating MCs. The naive union approach, which combines outputs from multiple prompts, achieves high recall but also high yield. In contrast, the individual prompts achieve a low recall (0.66-0.69), demonstrating the benefit of combining outputs from multiple prompts to

improve MC coverage. Furthermore, our clustering approach, which groups semantically similar MCs together, achieves a good balance of recall (0.83-0.88) and low yield (1638 - 406 words). Since vectors closer to each other represent MCs similar in meaning, a lower delta (δ) value only groups MCs that are nearly identical—this creates many small clusters because even slight differences in meaning are preserved as separate groups (higher recall) but requires review of more MCs (higher yield). Conversely, a higher δ value allows moderately similar MCs to group together, resulting in fewer, larger clusters (lower yield) but with a small drop in recall. Finally, we observe that the decomposition step, where the LLM is prompted to break down its output into simple statements, does not significantly affect the performance, suggesting it can help produce simple MCs without compromising performance. Overall, our two-stage approach of combining prompts and applying clustering offers an effective way to generate reliable, concise MCs using LLMs.