Automated main concept generation for narrative discourse assessment in aphasia

Ankita Gupta[♠] **Marisa Hudspeth**[♠] **Polly Stokes**[♡]
Jacquie Kurland[♡] **Brendan O'Connor**[♠]

[♠]College of Information and Computer Sciences
[♡]Department of Speech, Language & Hearing Sciences
University of Massachusetts Amherst
{ankitagupta,mhudspeth,brenocon}@cs.umass.edu
{pestokes,jacquie}@umass.edu

Abstract

We present an interesting application of narrative understanding in the clinical assessment of aphasia, where *story retelling* tasks are used to evaluate a patient's communication abilities. This clinical setting provides a framework to help operationalize narrative discourse analysis and an application-focused evaluation method for narrative understanding systems. In particular, we highlight the use of *main concepts* (MCs)—a list of statements that capture a story's gist—for aphasic discourse analysis. We then propose automatically generating MCs from novel stories, which experts can edit manually, thus enabling wider adaptation of current assessment tools. We further develop a *prompt ensemble* method using large language models (LLMs) to automatically generate MCs for a novel story. We evaluate our method on an existing narrative summarization dataset to establish its intrinsic validity. We further apply it to a set of stories that have been annotated with MCs through extensive analysis of retells from non-aphasic and aphasic participants (Kurland et al., 2021, 2025). Our results show that our proposed method can generate most of the gold-standard MCs for stories from this dataset. Finally, we release this dataset of stories with annotated MCs to spur more research in this area.[1]

1 Introduction

Narratives are an essential tool for communication and understanding. Several advances have been made towards developing theoretical (Piper et al., 2021; Genette, 1983) and computational methods (Chambers and Jurafsky, 2008; Goyal et al., 2010) for understanding narratives. However, evaluating the effectiveness of the obtained narratives has remained challenging owing to the subjective nature of the task and the difficulty of validating them in real-world applications where they serve specific communication purposes (Santana et al., 2023). For instance, while prior work has highlighted various downstream applications that can benefit from automated narrative understanding systems (e.g., developing intelligent tutoring systems (Halpin et al., 2004; Passonneau et al., 2007), predicting mental well-being (Adler et al., 2016)) the evaluation of such systems has been difficult owing to the limited availability of application-specific frameworks and datasets.

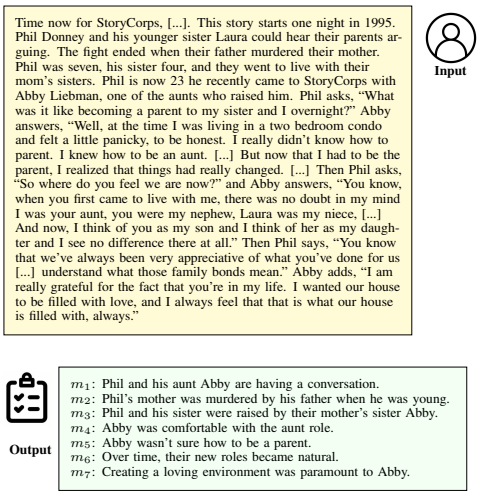

Figure 1: An example of the main concept generation task, showing an audio story's transcript and its gold-standard main concepts which have been used for clinical assessment of aphasia (Kurland et al., 2021).

In this work, we seek to bring to attention the prevalent practices of understanding narrative discourse when clinically assessing people with language impairments. These practices offer both a way to operationalize the task of identifying essential content in a narrative and a validated framework for evaluating narrative understanding systems through real-world healthcare-related communication tasks, which we hope can benefit the current research on computational narratology and NLP more broadly.

In particular, we focus on aphasia—a language impairment caused by stroke or other brain injury that affects a person's ability to express and understand written and spoken language. Clinical as-

[1] https://github.com/slanglab/aphasia

Findings of the Association for Computational Linguistics: ACL 2025, pages 24437–24451

sessment of functional communication in aphasia often relies on *story-retelling tasks*, where clinicians systematically analyze if patients can understand and convey essential elements of a story they just watched/heard (Kurland et al., 2021). These clinical assessments use a checklist of *main concepts*—a set of statements that capture a story's gist, with each statement consisting of one main verb and its subject, object, modifiers, and subordinate clauses if appropriate—which are empirically derived through extensive analysis of hundreds of retellings from healthy participants (Kurland et al., 2021; Richardson and Dalton, 2016, 2020) and have been used to assess patients with aphasia (Kurland et al., 2025). Figure 1 shows an example of such a story and its gold-standard main concepts.

While the current assessment tools have made advances by using manually pre-determined *main concepts* for a small set of stories (Kurland et al., 2021, 2025), the manual effort entailed severely limits the adaptation of these tools to include new stories. This constraint has significant clinical implications: patients may become familiar with the stories over repeated assessments, available stories may not be culturally relevant to a patient, and clinicians lack the flexibility to introduce new patient-centered assessment material that could provide more accurate insights into patient abilities (Thiessen and Brown, 2021).

Recent advances in NLP have shown great promise in automating various tasks involved in narrative understanding, including character identification (Brahman et al., 2021), inferring latent personas (Bamman et al., 2013), their relationships (Iyyer et al., 2016) and emotions (Brahman and Chaturvedi, 2020), event chains (Chambers and Jurafsky, 2008), summarizing (Zhao et al., 2022; Kryscinski et al., 2022) and identifying overall narrative structures (Boyd et al., 2020), *inter alia*. These advances present an opportunity to automate the *main concept* generation task for novel and patient-centric stories, enabling the rapid development of new assessment materials.

Our major contributions include:

1. We present an interesting application of narrative understanding in the clinical setting, where *story retelling* is used for assessing the communication abilities of patients with aphasia using *main concept* analysis.

2. We introduce the task of *main concept generation* for a given story, which involves generating a list of statements that convey the gist of the story. Unlike free-form text summaries often used in prior work on narrative summarization, our semi-structured output format aims to facilitate manual reviewing/editing and selection by experts and to be later used as a checklist for clinicians when evaluating patient retells.

3. We next use large language models (LLMs) to generate MCs for a given story. In particular, we develop a *prompt ensemble method* to generate MCs, which not only helps alleviate prompt sensitivity issues (different prompts generating different sets of MCs) but also helps generate a comprehensive and concise list of MCs. We evaluate our proposed method on an existing narrative summarization dataset (Zhao et al., 2022) adapted to our task, to establish our method's intrinsic validity.

4. Finally, we apply our method to a set of stories used in clinical aphasia assessment, where gold-standard MCs were collected from hundreds of non-aphasic participants (Kurland et al., 2021) and have been demonstrated useful for the assessment of hundreds of aphasia patients, showing a strong correlation with other clinical measures (Kurland et al., 2024, 2025). We find our method can generate most of the gold-standard main concepts for these stories. We further provide qualitative analysis of model errors to identify promising directions for future research. Finally, we release this dataset to facilitate further development and evaluation of narrative understanding systems.

Overall, our approach to improving methods for the assessment of narrative communication can have broad applications beyond aphasia, including communication-impaired populations across the lifespan (e.g., traumatic brain injury, dementia, mild cognitive impairments). More broadly, our method for generating main concepts from narratives can also be applied in other applications such as assessing student's story rewrites/retells for educational purposes (Halpin et al., 2004; Passonneau et al., 2007) or summarizing litigant narratives for attorneys (Branting et al., 2023).

2 Narrative discourse analysis in clinical assessment of aphasia

We next provide an overview of the current practices of assessing narrative discourse in the clinical assessment of aphasia.

Assessing language and communication impairment. Chronic impairments in communication can have devastating impacts on a person's identity, quality of life, and social functioning (Ayerbe et al., 2013). As such, aphasia contributes to a high incidence of post-stroke clinical depression (Cruice et al., 2010), highlighting the importance of assessment and monitoring tools.

Traditional assessment methods rely on word-level tests (Goodglass et al., 2001; Kertesz, 2006) or picture-based speech samples (Pashek and Tompkins, 2002). However, these tests cannot capture how well patients communicate in everyday situations, leading to the adoption of a life participation approach to aphasia (Chapey et al., 2000) focusing on more naturalistic assessment approaches. In particular, recent tools (e.g., Kurland et al., 2021, Ramsberger and Rende, 2002, Carragher et al., 2015), have proposed the use of *story-retelling* tasks, where patients are asked to convey a story they have just watched/heard to their conversation partners (Figure 2). This dyadic setting helps better reflect real-world communication scenarios (Kurland et al., 2021; Carragher et al., 2024).

Main concept analysis. A common method for analyzing aphasic discourse is main concept analysis (MCA) (Nicholas and Brookshire, 1995), a measure of how well a speaker communicates the essential concepts or gist of a discourse. Clinical researchers have developed standardized checklists of main concepts based on large samples of aphasic and non-aphasic subjects describing the same content (Richardson and Dalton, 2016, 2020; Kurland et al., 2021). As shown in Figure 2, during an assessment, clinicians or an automated system (Kurland et al., 2025) compare a patient's story retelling against these checklists, rating each main concept from 0 (not present) to 3 (accurate and complete). As a result, MCA has become a *clinician-friendly* means of compensating for the time- and resource-intensive burden of analyzing discourse in aphasia (Kim et al., 2019, 2021).

Process of developing main concept checklists. The development of MC checklists follows a rigorous process designed to ensure reliability and comprehensiveness (Richardson and Dalton, 2016, 2020; Kurland et al., 2021). We detail this process through the example of the Brief Assessment of Transactional Success (BATS) assessment tool (Kurland et al., 2021), which we also use for our experiments in §5.4.

The BATS provides 8 short video/audio clips as stories (mean duration = 2.86 minutes; SD = 0.37 minutes; Table 1). These stories include short interviews and autobiographical news clips, and many evoke emotional responses, including humor, sadness, awe, and inspiration. The process of obtaining the MCs for each story involved analyzing 768 story retells from 96 non-aphasic participants who were asked to retell each story "in as much detail as [you] can remember." Human raters then compiled a list of candidate MCs — statements containing essential elements like one main verb with its subject, object, and modifiers — and manually coded them for presence, accuracy, and completeness on a 0-3 scale in each retelling. Candidate MCs that were accurately and completely produced by at least 33% of participants were established as MCs, forming a gold standard for assessment.

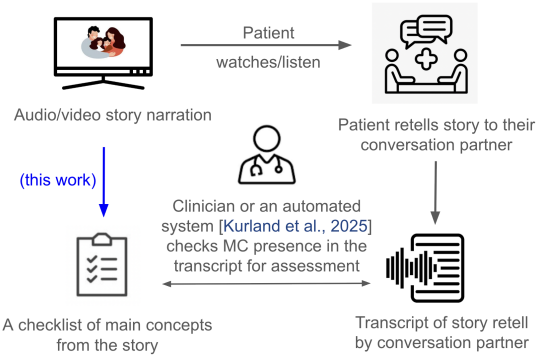

Figure 2: Work flow of aphasia assessment using *story retelling* task. The patient with aphasia retells a story they just watched/heard to their conversation partner. The clinician compares the transcript of the partner's story retelling with a list of pre-determined main concepts (MCs) from the original story for assessing the patient's functional communication. While prior work relies on manually curated MCs for a small set of stories, in this work we aim to automate the MC generation process, thereby enabling adaptation of existing assessment tools to include novel personalized stories.

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

Table 1: Descriptive information on BATS stimuli. VS: visually supported biographical video; SD: speech-dependent audio clip with a single still photo for visual support. VS clips are from the *PBS Brief but Spectacular* series and SD are from *NPR StoryCorps* series.

The need for automated main concept generation. Current assessment tools present several limitations. Existing stimuli often depict family scenes with outdated gender roles that few contem-

porary people with aphasia can relate to (Dalton et al., 2024) and thus may not accurately reflect their true communication abilities in everyday situations (Thiessen and Brown, 2021). Labor-intensive manual curation of MCs limit the tool's adaptation to new stories. Thus, an automated approach to MC generation could enable the adaptation of assessment tools to novel stimuli personalized for patients with different cultural backgrounds, potentially revolutionizing the intervention landscape of aphasia and other language impairments.

3 Main concept generation task

Given the input text of a story (T), which can be either an audio transcript or obtained from a written source, the task involves generating a list of main concepts $M = \{m_1, m_2....m_n\}$. Each concept m_i is a statement that is relevant to the story and consists of one main verb and its subject, object, modifiers, and subordinate clauses if appropriate. This definition of an MC follows the precedent established in the aphasia and speech/language disorders literature. The same definition was provided to human annotators when creating the original gold-standard MCs used in clinical practice (Kurland et al., 2021; Richardson and Dalton, 2016, 2020). Prior clinical research has demonstrated that MCs defined in this manner provide meaningful and reliable measures for clinical assessment applications (Kurland et al., 2025). The number of main concepts varies across stories, depending on their content. The set of main concepts together conveys the gist of the story.[2]

Figure 1 shows an example drawn from BATS, showing an audio story's transcript and its annotated main concepts. The story captures a conversation between a nephew and his aunt, reflecting on how the aunt became a mother figure to him and his sister after a tragedy, exploring their evolving family bonds and shared journey. This story has seven annotated main concepts that capture the summary of this conversation.

4 Prompt ensemble method for MC generation

Step 1: Generating MCs. We prompt an LM to generate main concepts in a zero-shot setting, without providing any examples. Since LM outputs can vary based on prompt wording, we experiment with multiple prompts.

Prior work demonstrates that using names of pedagogically popular theories can be an effective way to design prompts (Gupta et al., 2024). Thus, we prompt LM with names of theories used for understanding story grammar (Mandler and Johnson, 1977) or five-finger story-retelling strategies which are often used to teach students how to analyze and summarize narratives effectively (Baumann and Bergeron, 1993). We also prompt with 5W's and 5W's1H framework often used by journalists to understand key story aspects (Kroll, 2018). In each prompt, we provide the theory/concept definition and instruct the LM to generate MCs as per the theory/concept. Additionally, as baselines, we prompt LM with simpler prompts wherein we directly ask the LLM to help understand the story or explain the plot points of the story.[3] Detailed prompts are provided in Appendix A.1.

The raw LM response generated by the LM can often contain lengthy sentences combining multiple concepts. Since our objective is to generate MCs that are amenable to manual review, we further prompt the LLM to decompose its outputs into simpler de-contextualized statements. This step also allows us to generate MCs that structurally match the gold-standard MCs, further facilitating clustering in later steps. We use the few-shot prompt by (Tang et al., 2024) for this step, asking the LM to decompose each sentence in the raw LM response into simpler statements.[4]

Step 2: Union of MCs generated by different approaches. Prior work often selects the best-performing prompt, but we found this insufficient as different prompts and samples obtained via the same prompt can capture different story aspects and hence generate different main concepts (§5.4).

Previous studies have explored using prompt sensitivity to improve downstream task performance by aggregating word-level LLM outputs (Wang et al., 2023; Cai and O'Connor, 2023). Building on these efforts, we perform a set union of decomposed and de-contextualized statements obtained from the previous step. In particular, we take a

[2]Current clinical assessment tools evaluate aphasic patients' retells without considering the sequential order among MCs (Kurland et al., 2021). Following this practice, we evaluate automated method outputs without considering MC order. Future work could explore order-aware evaluation methods.

[3]We also experimented with chain-of-thought prompting (Kojima et al., 2022) and anecdotally found that it did not result in any performance gains compared to the prompts considered in our work. Furthermore, Sprague et al. (2024) also reports that COT prompting is often more useful for tasks involving math and symbolic reasoning than language understanding and information extraction tasks such as ours.

[4]Before decomposition, we remove the header and introductory text from LM's raw response through string matching against a manually compiled list of common expressions across all outputs (e.g., "here are the MCs").

union of all statements obtained from the three best-performing prompts and their samples to improve coverage of gold-standard MCs.

Step 3: Semantic deduplication. While a naive union of outputs from different prompts can help improve the coverage of gold-standard MCs, 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 number of words in the MC list—making manual review cumbersome.

One approach to address this issue would be to use an LLM for deduplication, by giving it the union list of MCs and prompting it for a deduplicated MC list. However, this approach faces several limitations. First, it is constrained by the length of the context-window of an LLM; for long stories with a large number of MCs, the input list of MCs to deduplicate across different prompts can quickly become large and may not fit in the context window. Secondly, even as LLMs continue to support longer context windows, they still struggle to actually utilize the entire context. For instance, the recent needle in a haystack analysis (Kamradt, 2023), which tests the in-context retrieval ability of long context LLMs, suggests that both open and closed-sourced LLMs struggle with utilizing long context. Furthermore, LLMs have also been shown to perform poorly on tasks requiring long context understanding, such as book-length narrative summarization (Chang et al., 2024; Kim et al., 2024). For deduplication, such capabilities are all the more necessary.

Instead, we use a clustering-based approach that groups MCs that are similar in meaning and selects one representative MC from each cluster. We first convert each MC into an embedding using Sentence-BERT (Reimers and Gurevych, 2019). We then L2 normalize the embedding vectors, followed by clustering.

We cluster with the DP-means algorithm (Kulis and Jordan, 2012),[5] which Hanley et al. (2025) found useful for embedding-based semantic clustering. In DP-means, the granularity of clusters is controlled by a hyperparameter $\delta \geq 0$, the maximum squared distance between a data point and its cluster's centroid; the model chooses the total number of clusters in response to that constraint. A small δ gives a large number of spatially small clusters, each with high internal semantic similarity; and conversely for large δ. We report results in terms of δ, which is roughly twice of the familiar cosine distance.[6] Finally, to obtain one representative MC for each cluster, we select the embedding closest to the centroid and use its text as the representative MC. The representative MCs form our final MC list.

5 Experiments

5.1 Experimental setup

Language models. We consider one of the recent state-of-the-art open-weight LLMs, `Llama-3.3-70B` for all our main experiments.[7] We choose open-weight models in contrast to proprietary models, as they can be downloaded and used locally, providing more flexibility for customization (e.g., adapting model parameters for specific tasks), interpretability, cost efficiency, and improved privacy, as they do not require sending sensitive data to external servers, which is of great concern for clinical applications. Our choice also aligns with recent efforts on using open-source models to automatically assess MCs in patient retells (Kurland et al., 2025), providing avenues for developing an ecosystem for clinical assessment of aphasia using open-weight models. We use a temperature of 0.67 to generate 5 samples per prompt and report average performance over samples for each prompt. To examine whether the prompt sensitivity issue is model-specific, we consider another alternative LM, `DeepSeek-V3`,[8] though observe similar sensitivity issues.

5.2 Evaluation metrics

We measure the quality of generated MCs using the following two metrics:

Recall. First, we measure the quality of LLM-generated MCs by comparing them to the gold-standard MCs. In particular, we compute *recall*, which measures the fraction of gold-standard MCs present in the LLM-generated MCs. We consider a generated MC as acceptable if it is supported by the human-curated gold-standard MCs. Since the human-curated MCs already adhere to the intended structure requirements of an MC, our recall-based evaluation also accounts for structural assessment of the generated MCs.

Prior work has used `GPT4` to evaluate various NLP tasks (Gu et al., 2025; Liu et al., 2023; Min

[5] Via Dinari and Freifeld (2022)'s parallel implementation of DP-means.

[6] For unit norm sentence embeddings, their squared distance is $\|\mathbf{a}-\mathbf{b}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2\langle \mathbf{a}, \mathbf{b}\rangle = 2(1 - \cos(\mathbf{a}, \mathbf{b}))$. This statement does not precisely apply to δ since centroids generally do not have a unit norm, due to convexity of the unit ball.

[7] Accessed via `https://www.together.ai/`

[8] Accessed via `https://hyperbolic.xyz/`

Dataset	Average number of tokens		Compression Ratio
	Documents	Summaries/MCs	
BATS	481	75	0.16
NARRASUM	895	213	0.24

Table 2: Statistics of evaluation datasets. The compression ratio denotes the ratio of summary length to the document length.

et al., 2023). Building on these efforts, we also use it as an automatic evaluator to compute recall. Our evaluation focuses on recall, as we expect clinicians to review and refine the generated main concept lists rather than relying on them as final outputs.

In particular, given the LLM-generated and gold-standard MCs, we ask `GPT-4o` to identify which of the gold-standard MCs are supported by the LLM-generated MCs. If a gold-standard MC is supported, we ask GPT-4 to provide the span of text from the LLM-generated MCs that support the answer; otherwise, explain why it is not supported. The detailed prompt is provided in the Appendix A.2. For this evaluation, we prompt `GPT-4o` in a zero-shot setting at a temperature of 0.

Yield. Overly verbose or long MC lists can be difficult for manual review. Thus, we additionally measure the conciseness of LLM-generated MCs. One way to measure verboseness can be to count the number of generated MCs; however, a small list can contain lengthy statements. Thus, we calculate *yield* as the number of tokens summed over all MCs in the generated list. A lower *yield* (fewer tokens) is desirable as it indicates concise MCs, making them convenient for humans post-editing, sharing with others, or for clinical use.

5.3 Evaluation datasets

BATS dataset. To evaluate our method, we use the collection of stories and their MCs collected as part of the BATS study (§2). The MCs provided for each story in BATS have been manually curated and validated over multiple aphasic and non-aphasic participants, offering a robust benchmark for evaluating our automated MC generation method while ensuring clinical relevance. We consider all stories and their annotated MCs as mentioned in Table 1.

`NARRASUM` dataset (Zhao et al., 2022). Since the number of stories annotated with gold-standard MCs is small in the BATS dataset, we additionally evaluate our proposed MC generation method using the `NARRASUM` dataset to establish our method's validity. This dataset provides an evaluation dataset of 100 story-summary pairs sourced from plot descriptions of movie/TV shows. This dataset is particularly relevant to our task for several reasons: a) Similar to the BATS dataset, this dataset also aims to support narrative summarization which involves producing a distilled version of a narrative to describe its most salient events and characters, b) Similar to the stories used in the BATS dataset, the documents in `NARRASUM` dataset are narrative texts and thus naturally include the story arcs capturing the sequence of events and major characters with their profiles (e.g., personalities, roles, and interpersonal relationships), c) Similar to MCs in BATS, the `NARRASUM` summaries include main events that significantly impact the plot development, motivations of the main characters, their consequences, and the ending of the main event, d) Finally, the compression ratio of both datasets is in a comparable range as shown in Table 2.

However, the summaries in `NARRASUM` are written as free-text paragraphs, while our task requires a list of MCs. To adapt the `NARRASUM` dataset, we first segment each summary paragraph into sentences. Since these summary sentences are significantly longer than typical MCs in the BATS dataset (`NARRASUM`: 20.28 tokens vs BATS: 8.09 tokens), we further decompose each sentence into concise MCs using the decomposition step as used in our MC generation step (§4). We use `GPT-4o` for these decompositions using the few-shot prompt as used in §4. Each story in `NARRASUM` dataset has an average of 28 gold MCs, a minimum of 5 gold MCs, and a maximum of 120 gold MCs.

5.4 Results on the `NARRASUM` dataset.

Table 3 shows the recall obtained using `Llama-3.3-70B` when prompted with different prompts. The top three performing prompts are `help me understand the story`, `5W's retell strategy`, and the `five-finger retell strategy`, all with a recall of 0.46. Additionally, to validate our evaluation method, we test simple baseline prompts asking LM to generate only one or two plot points in the story. As expected, these prompts achieve lower recall (≤ 0.30), since these prompts require an LM to generate only a subset of the story's main concepts.

Different sets of MCs are generated by different prompts and samples. Given that the three best-performing prompts achieve similar recall scores of 0.46, we further investigate whether they are identifying the same or different subsets of gold MCs. To understand this, we calculate the Jaccard distance (J_d) between sets of gold MCs generated by pairs of prompts. As shown below, (J_d) quantifies the

Prompt	NARRASUM		BATS	
	Recall	Yield	Recall	Yield
Five-finger retell strategy	0.46±0.02	247.87±7.39	0.68±0.04	754.50±13.27
7-unit-story grammar	0.43±0.04	245.85±6.25	0.59±0.03	1005.60±16.61
5W's 1H retell strategy	0.44±0.02	308.30±7.88	0.47±0.05	223.22±17.30
5W's retell strategy	0.46±0.02	232.31±6.77	0.50±0.04	177.40±13.85
Help me understand the story	0.46±0.04	307.91±6.64	0.66±0.10	1448.00±10.74
Identify one plot point	0.17±0.01	32.80±1.17	0.28±0.05	37.75±05.71
Identify two plot points	0.30±0.02	79.53±2.04	0.39±0.05	78.03±06.68
Theoretical Union	0.87	3953.10	0.92	3109.10
Empirical Union	0.64	3953.10	0.89	3109.10
Prompt ensemble (ours)	0.53	504.00	0.87	633.90

Table 3: Recall and yield scores on the NARRASUM and the *BATS* dataset for different prompting strategies and our proposed method. Confidence intervals are calculated over multiple samples of a single prompt. Unions are taken over the three best-performing prompts.

diversity among two sets (S_1, S_2) by comparing the size of their intersection to the size of their union.

$$J_d(S_1, S_2) = \left(1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}\right)$$

A J_d close to 0 means the sets are very similar, while a J_d close to 1 means the sets are very different.

We examine the diversity among sets of gold MCs generated by different prompts or different samples of the same prompt. We consider the three best-performing prompts for this analysis. For each story in the NARRASUM dataset, we calculate the J_d between sets generated by different pairs of prompts/samples. We then compute the average J_d across all stories to quantify the diversity of sets generated by a pair of prompts/samples.

Comparison Type	Prompt 1	Prompt 2	Jaccard Distance
Pair of samples	Help me understand	Help me understand	0.46
	5W's retell	5W's retell	0.46
	Five-finger retell	Five-finger retell	0.42
Pair of prompts	Help me understand	5W's retell	0.28
	Help me understand	Five-finger retell	0.28
	5W's retell	Five-finger retell	0.30

Table 4: Jaccard distances among sets of gold MCs generated by different prompts or samples of the same prompt on the NARRASUM dataset.

Table 4 shows high diversity ($J_d : 0.42 - 0.46$) among sets generated by samples of the same prompt. Furthermore, when comparing sets generated by different prompts, despite each prompt having comparable recall scores, the J_d of $0.28 - 0.30$ shows that different prompts capture distinct story aspects, leading to different sets of gold MCs. This observation highlights that while different prompts may identify a similar number of gold MCs, they possibly capture different narrative elements in a story, resulting in the generation of diverse sets of MCs. Overall, the high diversity of sets across both samples and prompts suggests that choosing the best-performing prompt will be insufficient to comprehensively extract all MCs from a story. We also experimented with DeepSeek-V3 and found similar prompt sensitivity issues (sample J_d: 0.29-0.36, prompt J_d: 0.30-0.34).

Union of MCs generated by different prompts helps improve recall but also increases the yield. Given the diverse sets generated by different prompts and samples, we explore whether taking a set ensemble can help improve recall. We examine ensembles of sets generated by samples of the same prompt, different prompts, and both.

When evaluating these union sets, we compute two types of recalls: *theoretical union* recall and *empirical union* recall. In the *theoretical* approach, we first identify the set of gold MCs identified by each prompt/sample. Then, we take the union of these successful matches across all prompts and samples—if any prompt/sample identifies a gold MC, we include that gold MC in the final set. This method helps us understand how different prompts/samples complement each other in capturing different gold-standard MCs. In the *empirical* approach, we first combine all the generated MCs from all prompts and samples into a single set. We then evaluate this set against the gold standard MCs to determine which gold MCs are present. This approach provides a more straightforward evaluation of the combined output but needs longer LLM context length, can be expensive due to lengthy input prompts, and is more error-prone as the LLM must process a large number of MCs simultaneously.

Table 3 shows the *theoretical union* recall and yield obtained using Llama-3.3-70B when taking a union of the MCs generated from all samples across all prompts. We observe that this naive ensemble increases recall to 0.87 beyond what any individual prompt achieves $(0.43 - 0.46)$. This improvement in recall comes from both samples and prompts—combining samples from individual prompts boosts *theoretical union* recall to $0.69 - 0.73$, and combining pairs of prompts further increases it to $0.82 - 0.83$. However, it also results in a very high yield (3953.10).

We also observe that the *empirical union* recall is 0.64, higher than the recall from the three prompts, though much lower than the *theoretical union* recall. On qualitative examination, we find that the *empirical union* recall is lower than the *theoretical union* recall often for stories with a very large number of gold MCs (≥ 40), where the lengthy prompts make evaluation more challenging. Given our BATS use case involves stimuli with fewer MCs (≤ 12), we retain the current evaluator leaving improvements for longer prompts as future work.

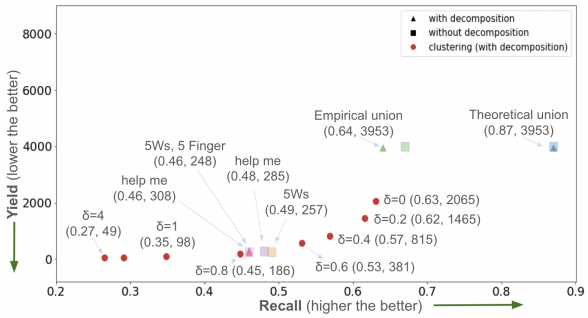

Figure 3: Recall v/s yield tradeoff on NARRASUM dataset.

Semantic de-duplication can help control for recall versus yield tradeoff. Figure 3 shows the recall versus yield tradeoff using our set valued ensemble approach, where we take a union of MCs generated by different prompts/samples followed by semantic de-duplication using clustering.

Our approach achieves a balance of higher recall (0.51-0.62) with a lower yield (381-1465 words). We observe that at $\delta = 0$, only exact matches are clustered and hence the recall is comparable to the *empirical union*, with a slightly lower yield (1465). A lower 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 require review of more MCs (higher yield). Conversely, higher δ values > 1 allow many MCs to be grouped together, resulting in fewer, larger clusters (lower yield: 49-98), but substantially reducing recall to 0.27-0.35. Finally, selecting a δ value that gives a yield comparable to the yield of the individual prompts (232-308), we observe a recall of 0.53 with a yield of 381 at $\delta = 0.6$. Overall, our results show that a reliable and concise list of MCs can be generated using our prompt ensemble method.

Sensitivity analyses: Impact of MC decomposition step. We further compare the recall obtained before and after the decomposition step (where the LLM is prompted to break down its output into simple statements), for each of the three-best performing prompts. As shown in Figure 3, the decomposition step does not significantly affect the performance, suggesting it can help produce simple MCs without compromising performance.

5.5 Results on BATS dataset

Having established the intrinsic validity of our prompt ensemble method, we next apply it to generate MCs for BATS stories and compare them with BATS' gold-standard MCs.

Recall versus yield tradeoff. Table 3 shows that the `five-finger retell strategy` prompt achieves the highest recall (0.68), followed by the `7-unit story grammar` (0.59) and the general `help me understand the story` prompt (0.66). We use these three prompts for ensemble approach.

Figure 4 shows the recall versus yield tradeoff obtained after our semantic deduplication step. Union over outputs from different prompts/samples obtains a high recall (0.89), 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. The semantic de-duplication step further helps reduce yield with only a small drop in recall. For instance, selecting δ with a yield comparable to individual prompts ($731 - 754$), we observe that at $\delta = 0.4$, the semantic de-duplication step using clustering results in a yield of 634 words, much smaller than the naive union (3109 words), while maintaining a high recall of 0.87. Overall, our results show that our prompt ensemble approach can be used to generate a comprehensive yet concise list of MCs.

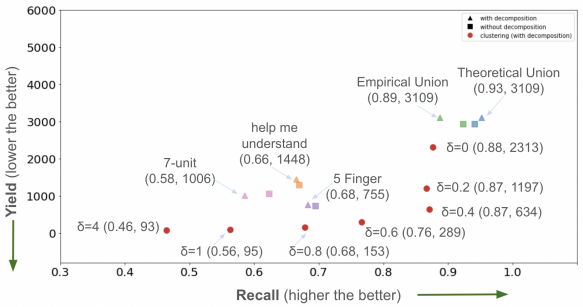

Figure 4: Recall vs. yield tradeoff on the BATS dataset.

Qualitative analysis. We conduct a qualitative analysis to understand the model's mistakes in identifying gold-standard MCs for a story. LMs often miss MCs that describe character feelings (e.g., 'Abby was comfortable with the aunt role') and desires (e.g., 'The mother wants her son to be selfish'), or observations about the environment (e.g., 'The passengers were silent', 'The church was packed'), across all prompt variants. Developing methods that can identify these specific narrative components is an interesting avenue for future work.

6 Related work in NLP

Our work is related to the following research areas:

Narrative understanding. Several advances have been made towards developing theoretical

frameworks to support computational narrative understanding (Piper et al., 2021; Piper, 2023), outlining essential narrative components, such as characters, their internal states (Zhu et al., 2023), events. Prior work has also developed computational methods to extract and represent these narrative components as highly structured representations (Santana et al., 2023), including plot points (Goyal et al., 2010), story graphs (Valls-Vargas et al., 2017), entity-based narrative graphs (Lee et al., 2021), intention graphs (Lukin et al., 2016). However, using these representations can be difficult for clinicians to use. In contrast, our work involves generating a list of MCs that captures a broader range of narrative elements, including explicit events and implicit character states. A list of MCs is also convenient for humans post-editing and clinical use.

Narrative summarization. Our task is also related to narrative summarization, where each MC also aims to capture key story elements such as characters, their relationships, and major events. However, unlike free-form text output common in prior work (Zhao et al., 2022; Kryscinski et al., 2022), our task aims to generate a list of statements capturing the gist of the story. MCs can also be viewed as narrative content units, similar to the concept of summarization content units studied in Nenkova and Passonneau (2004). Our task can also be viewed as a form of key point analysis task (Bar-Haim et al., 2020a; Egan et al., 2016; Bar-Haim et al., 2020b) which aims to select and summarize the most important points from an input narrative.

7 Conclusion

We present an application of narrative understanding in the clinical assessment of aphasia using story retelling, which we hope can benefit current research in developing and evaluating narrative understanding systems. We further propose the task of main concept generation to enable the adaptation of current assessment tools to novel and patient-centric stories. We also present an LLM-based prompt ensemble method for automatically generating MCs from stories. Our experimental results on both the BATS clinical dataset and the NARRASUM dataset demonstrate that our proposed method can successfully generate a concise list of MCs that match manually curated gold-standard MCs. More broadly, our main concept generation approach can also benefit broader applications such as assessing functional communication across diverse clinical populations and in educational contexts like intelligent tutoring systems.

8 Limitations

We list some of the limitations of our study, which we hope will be useful for researchers and practitioners when interpreting our analysis.

1. Our semantic de-duplication approach relies on clustering, but there is room for improvement. For instance, Pham et al. (2024) used LLMs to generate topics from a sample of documents and to merge repeated entries from a given list of topics. Building on this effort, LLMs can be used to obtain better cluster representatives for each cluster, instead of choosing the MC closest to the cluster centroid. Similarly, as LLMs continue to support longer context windows, future work could explore their ability to directly de-duplicate lengthy lists of MCs through prompting.

2. In this work, we have considered several different prompts motivated by educational material. However, alternative prompting strategies are possible. For example, chain-of-density prompting (Adams et al., 2023), which focuses on entity-focused summarization by including all mentioned characters and entities in the narrative, could be used for generating MCs. Since our prompt ensemble approach is not constrained to the specific prompts considered in our work, such alternative prompts can be easily integrated with our approach. These prompts may also capture different narrative aspects and potentially improve overall performance.

3. Our recall evaluator is based on LLM-prompting and struggles with longer prompts. However, using LLMs for evaluation is an active area of research (Li et al., 2025), and improving recall evaluator via task-specific fine-tuning (Tang et al., 2024) or in-context learning can be interesting future work.

4. Our approach currently uses open-weight LLMs accessed via paid APIs, and thus can be expensive when analyzing a large number of stories. Future work can explore fine-tuning smaller models or examining the capabilities of smaller models on this task, to reduce the cost considerations.

9 Ethics Statement

Our work is in line with the ACL Ethics Policy. The text and appendix outline all the models, datasets, and evaluation methodologies used in this research.

All evaluation datasets used in this research are publicly available or used with the appropriate consent. All the human subjects' details of the BATS dataset are described in the original dataset papers by Kurland et al. (2021, 2025). This research was conducted in collaboration with speech-language pathologists with expertise in aphasia assessment. No private or patient-specific information is included in our evaluations or dataset.

Acknowledgements

We would like to thank the anonymous reviewers for their time and valuable feedback. We are grateful to the UMass NLP group for several useful discussions during the course of the project. This material is based upon work supported by a UMass Interdisciplinary Research Grant, National Science Foundation award 1845576, the National Institute on Deafness and Other Communication Disorders of the National Institutes of Health award R21DC020265, and an IBM Ph.D. Fellowship award to AG. The content is solely the responsibility of the authors and does not necessarily represent the views of any sponsor.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics.

Jonathan M Adler, Jennifer Lodi-Smith, Frederick L Philippe, and Iliane Houle. 2016. The incremental validity of narrative identity in predicting well-being: A review of the field and recommendations for the future. *Personality and Social Psychology Review*, 20(2):142–175.

Luis Ayerbe, Salma Ayis, Charles D. A. Wolfe, and Anthony G. Rudd. 2013. Natural history, predictors and outcomes of depression after stroke: systematic review and meta-analysis. *British Journal of Psychiatry*, 202(1):14–21.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020a. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Roy Bar-Haim, Yoav Kantor, Lilach Eden, Roni Friedman, Dan Lahav, and Noam Slonim. 2020b. Quantitative argument summarization and beyond: Cross-domain key point analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 39–49, Online. Association for Computational Linguistics.

James F. Baumann and Bette S. Bergeron. 1993. Story map instruction using children's literature: Effects on first graders' comprehension of central narrative elements. *Journal of Reading Behavior*, 25(4):407–437.

Ryan L Boyd, Kate G Blackburn, and James W Pennebaker. 2020. The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, 6(32):eaba2196.

Faeze Brahman and Snigdha Chaturvedi. 2020. Modeling protagonist emotions for emotion-aware storytelling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.

Faeze Brahman, Meng Huang, Oyvind Tafjord, Chao Zhao, Mrinmaya Sachan, and Snigdha Chaturvedi. 2021. "Let your characters tell their story": A dataset

for character-centric narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1734–1752, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Karl Branting, Sarah McLeod, Bryant Park, and Karine Megerdoomian. 2023. Induction of narrative models for legal case elicitation. In *ASAIL@ICAIL*.

Erica Cai and Brendan O'Connor. 2023. A monte carlo language model pipeline for zero-shot sociopolitical event extraction. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Marcella Carragher, Zaneta Mok, Gillian Steel, Paul Conroy, Kathryn Pettigrove, Miranda L. Rose, and Leanne Togher. 2024. Towards efficient, ecological assessment of interaction: A scoping review of co-constructed communication. *International Journal of Language & Communication Disorders*, 59(3):831–875.

Marcella Carragher, Karen Sage, and Paul Conroy. 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.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*.

Roberta Chapey, Judith F Duchan, Roberta J Elman, Linda J Garcia, Aura Kagan, Jon G Lyon, and Nina Simmons Mackie. 2000. Life participation approach to aphasia: A statement of values for the future. *The ASHA leader*, 5(3):4–6.

Madeline Cruice, Linda Worrall, and Louise Hickson. 2010. Health-related quality of life in people with aphasia: Implications for fluency disorders quality of life research. *Journal of Fluency Disorders*, 35(3):173–189.

Sarah Grace Dalton, Mohammed AL Harbi, Shauna Berube, and H. Isabel Hubbard. 2024. Development of main concept and core lexicon checklists for the original and modern cookie theft stimuli. *Aphasiology*, 38(12):1975–1999.

Or Dinari and Oren Freifeld. 2022. Revisiting DP-means: Fast scalable algorithms via parallelism and delayed cluster creation. In *The 38th Conference on Uncertainty in Artificial Intelligence*.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

Gérard Genette. 1983. *Narrative Discourse: An Essay in Method*, volume 3. Cornell University Press.

Harold Goodglass, Edith Kaplan, and Barbara Barresi. 2001. *The assessment of aphasia and related disorders*, 3rd edition. Lippincott Williams & Wilkins, Austin.

Amit Goyal, Ellen Riloff, and Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 77–86, Cambridge, MA. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on LLM-as-a-Judge. *Preprint*, arXiv:2411.15594.

Ankita Gupta, Ethan Zuckerman, and Brendan O'Connor. 2024. Harnessing Toulmin's theory for zero-shot argument explication. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10259–10276, Bangkok, Thailand. Association for Computational Linguistics.

Harry Halpin, Johanna D. Moore, and Judy Robertson. 2004. Automatic analysis of plot for story rewriting. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 127–133, Barcelona, Spain. Association for Computational Linguistics.

Hans W. A. Hanley, Yingdan Lu, and Jennifer Pan. 2025. Across the firewall: Foreign media's role in shaping Chinese social media narratives on the Russo-Ukrainian war. *Proceedings of the National Academy of Sciences*, 122(1):e2420607122.

Mohit Iyyer, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber, and Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1534–1544, San Diego, California. Association for Computational Linguistics.

Gregory Kamradt. 2023. Needle In A Haystack - pressure testing LLMs. GitHub.

Andrew Kertesz. 2006. *Western Aphasia Battery – Revised*. Harcourt Assessment, Inc.

Hana Kim, Stephen Kintz, and Heather Harris Wright. 2021. Development of a measure of function word use in narrative discourse: Core lexicon analysis in

aphasia. *International journal of language & communication disorders*, 56(1):6–19.

Hana Kim, Stephen Kintz, Kristen Zelnosky, and Heather Harris Wright. 2019. Measuring word retrieval in narrative discourse: Core lexicon in aphasia. *International Journal of Language & Communication Disorders*, 54(1):62–78.

Yekyung Kim, Yapei Chang, Marzena Karpinska, Aparna Garimella, Varun Manjunatha, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. FABLES: Evaluating faithfulness and content selection in book-length summarization.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

John Kroll. 2018. Digging deeper into the 5 W's of journalism. *International Journalist's Network*.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brian Kulis and Michael I. Jordan. 2012. Revisiting k-means: new algorithms via bayesian nonparametrics. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, page 1131–1138, Madison, WI, USA. Omnipress.

Jacquie Kurland, Anna Liu, and Polly Stokes. 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(1):39–68.

Jacquie Kurland, Anna Liu, Vishnupriya Varadharaju, Polly Stokes, and Robert Cavanaugh. 2024. Reliability of the brief assessment of transactional success in communication in aphasia. *Aphasiology*, 0(0):1–22.

Jacquie Kurland, Vishnupriya Varadharaju, Anna Liu, Polly Stokes, Ankita Gupta, Marisa Hudspeth, and Brendan O'Connor. 2025. Large language models' ability to assess main concepts in story retelling: A proof-of-concept comparison of human versus machine ratings. *American Journal of Speech-Language Pathology*, pages 1–11.

I-Ta Lee, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Modeling human mental states with an entity-based narrative graph. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4916–4926, Online. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. *Preprint*, arXiv:2411.16594.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Stephanie Lukin, Kevin Bowden, Casey Barackman, and Marilyn Walker. 2016. PersonaBank: A corpus of personal narratives and their story intention graphs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1026–1033, Portorož, Slovenia. European Language Resources Association (ELRA).

Jean M. Mandler and Nancy S. Johnson. 1977. Remembrance of things parsed: Story structure and recall. *Cognitive Psychology*, 9(1):111–151.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Linda E Nicholas and Robert H Brookshire. 1995. Presence, completeness, and accuracy of main concepts in the connected speech of non-brain-damaged adults and adults with aphasia. *Journal of Speech, Language, and Hearing Research*, 38(1):145–156.

Gail V. Pashek and Connie A. Tompkins. 2002. Context and word class influences on lexical retrieval in aphasia. *Aphasiology*, 16(3):261–286.

Rebecca J. Passonneau, Adam Goodkind, and Elena T. Levy. 2007. Annotation of children's oral narrations: Modeling emergent narrative skills for computational applications. In *The Florida AI Research Society*.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long*

Papers), pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Andrew Piper. 2023. Computational narrative understanding: A big picture analysis. In *Proceedings of the Big Picture Workshop*, pages 28–39, Singapore. Association for Computational Linguistics.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative theory for computational narrative understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gail Ramsberger and Barbara Rende. 2002. Measuring transactional success in the conversation of people with aphasia. *Aphasiology*, 16(3):337–353.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Jessica D Richardson and Sarah Grace Dalton. 2016. Main concepts for three different discourse tasks in a large non-clinical sample. *Aphasiology*, 30(1):45–73.

Jessica D Richardson and Sarah Grace Hudspeth Dalton. 2020. Main concepts for two picture description tasks: An addition to Richardson and Dalton, 2016. *Aphasiology*, 34(1):119–136.

Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. 2023. A survey on narrative extraction from textual data. *Artificial Intelligence Review*, 56:8393–8435.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2024. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. *ICLR*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Amber Thiessen and Jessica Brown. 2021. Personalization of restorative and compensatory treatments for people with aphasia: A review of the evidence. *Topics in Language Disorders*, 41(3):269–281.

Josep Valls-Vargas, Jichen Zhu, and Santiago Ontañón. 2017. Towards automatically extracting story graphs from natural language stories. In *AAAI Workshops*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. NARRASUM: A large-scale dataset for abstractive narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 182–197, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lixing Zhu, Runcong Zhao, Lin Gui, and Yulan He. 2023. Are NLP models good at tracing thoughts: An overview of narrative understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10098–10121, Singapore. Association for Computational Linguistics.

A Appendix

A.1 MC generation prompts

Figures 5–11 show the different prompts used for generating MCs.

Suppose you are a story reteller. Use the following five finger retell strategy to identify the main concepts in the given story.

Five finger retell strategy:
1. Thumb - Somebody (Who are the characters?)
2. Pointer - Wanted (What do they want?)
3. Tall Finger - But (What was the problem?)
4. Ring Finger - So (What was the solution?)
5. Little Finger - Then (How did the story end?)

Story: {}
Please provide a list of main concepts mentioned in the above story using the five finger retell strategy.

Figure 5: Five finger retell prompt used for generating MCs.

Suppose you are a story reteller. Use the following 5 W's retell strategy to identify the main concepts in the given story.

Five W's retell strategy:
1. Who?
2. What?
3. When?
4. Where?
5. Why?

Story: {}
Please provide a list of main concepts mentioned in the above story using the 5 W's retell strategy.

Figure 6: 5W's prompt used for generating MCs.

Suppose you are a story reteller. Use the following 5 W's, 1 H retell strategy to identify the main concepts in the given story.

5 W's, 1 H retell strategy:
1. Who is the story about?
2. What did they do?
3. When did the action take place?
4. Where did the story happen?
5. Why did the main character do what s/he did?
6. How did the main character do what s/he did?

Story: {}
Please provide a list of main concepts mentioned in the above story using the 5 W's, 1 H retell strategy.

Figure 7: 5W's 1H prompt used for generating MCs.

A.2 Recall evaluation Prompts

Figure 12 provides the prompt used for recall computation using GPT4o.

Suppose you are a story reteller. Use the following 7-unit story grammar to identify the main concepts in the given story.

7-unit story grammar:
1. Setting
2. Characters in the story
3. Initiating Event [IE] – event that sets off the story's events – will cause the protagonist to respond in some way, evokes an immediate response
4. Internal Response [IR] – reaction of protagonist to the initiating event. It can be expressed in dialogue, e.g., oh no! expresses an internal response
5. Internal Plan [IP] of protagonist to deal with the IE
6. Attempt [ATT] to obtain the goal
7. Outcome or Consequence of the attempt

Story: {}
Please provide a list of main concepts mentioned in the above story using the 7-unit story grammar.

Figure 8: 7-unit story grammar prompt used for generating MCs.

Suppose you are a story-reteller, generate a list of main concepts to help me understand the story.

Story: {}

Figure 9: Help me understand prompt used for generating MCs.

Suppose you are a story reteller. Please identify one plot point in the given story in 30 words or less.

Story: {}

Figure 10: One plot point prompt used for generating MCs.

Suppose you are a story reteller. Please identify two plot points in the given story in 60 words or less.

Story: {}

Figure 11: Two plot point prompt used for generating MCs.

First, I will give you a list of main concepts. Then, I will give you a numbered list of candidate concepts. For each candidate concept, indicate whether it is supported by the list of main concepts. If it is supported, provide the span of text from the main concepts that support your answer; otherwise, provide an explanation of why it is not supported. Based on your explanation, predict “yes” or “no” for each candidate concept. Output your response by numbering the candidate concepts in the order they are presented.
Provide output in the following format exactly:

{{"MC": 1, "span": explanation, "is_supported": "yes" or "no"}},

{{"MC": 2, "span": explanation, "is_supported": "yes" or "no"}},
...

Main Concepts:
{generated_concepts_str}

Candidate Concepts:
{gold_concept_str}

Only provide the output in the specified format and nothing else (e.g., introductory texts, explanations, or reasons).

Figure 12: Recall evaluation prompt