Scalable Generation of Adult-Oriented Therapeutic Reading Texts for Russian Aphasia Rehabilitation

Anastasia Kolmogorova, Anastasia Margolina, Alina Telnova and Igor Ilchenko
National Research University Higher School of Economics, Russia
akolmogorova@hse.ru, avmargolina@edu.hse.ru, aatelnova@edu.hse.ru, igvladilchenko@gmail.com

Abstract

Texts are widely used in aphasia rehabilitation to support the recovery of comprehension and narrative planning. In routine practice, clinical impact depends strongly on patient motivation and on the availability of age-appropriate reading materials: adults are often offered child-oriented texts, which can be perceived as demeaning and may reduce engagement. We present a controllable generation pipeline for building a repository of Russian therapeutic reading texts for adult aphasia therapy. An anonymized repository with code and data is available at https://github.com/z00logist/aphasia-exercises-generation. The pipeline conditions each story on an explicit semantic triplet (12 topics $\times$ 10 archetypes $\times$ 11 objects) and enforces three clinically motivated complexity regimes (Basic/Intermediate/Advanced). Using batched prompting, we generate 1,296 unique stories. We evaluate the corpus with classical linguistic metrics, sentiment analysis, and a LLM-as-a-judge protocol (18 binary criteria); on a stratified sample of 198 stories, overall rubric compliance is 80.0%. Surface metrics show a monotonic increase in lexical and syntactic complexity across regimes, and comparisons against a clinical anchor set of 10 therapist-authored texts reveal that while generated texts match ground truth readability at the Basic level, standard clinical texts align closer to the Intermediate generated level in length and complexity. Judge-based analysis indicates near-perfect adherence to high-level narrative constraints but persistent limitations in fine-grained phonotactic control, which we manually confirm, motivating hybrid neuro-symbolic enforcement.

Keywords: aphasia therapy, controlled text generation, Russian, readability, LLM-as-a-judge, evaluation

1. Introduction

Texts are widely used by speech-language pathologists (SLPs) in speech therapy interventions for both children and adults presenting with various speech disorders or language impairments. Aphasia is one such systemic language disorder resulting from brain damage caused by traumatic injury, stroke, or surgical intervention (e.g., tumor resection) (Luria, 2008).

In speech therapy, texts serve as a tool for restoring different functional skills: for supporting the formation and execution of simple motor programs (e.g., during reading-while-listening tasks), for restoring connected text comprehension (e.g., tasks focused on understanding content), for vocabulary expansion, and for rehabilitating utterance planning skills (e.g., retelling) (Luria, 2008).

However, clinical practice indicates that a patient's motivation to read plays a crucial role in the effectiveness of text-based interventions in speech therapy (Webster et al., 2023). Motivation is shaped by several factors: whether the subject matter is engaging for the patient; whether linguistic complexity matches the patient's current functional level; and whether the text leaves a positive emotional impression after reading (Webster et al., 2023). Nevertheless, such texts are not read for entertainment, but rather as a means to achieve specific rehabilitation goals set by the SLP. The SLP, in turn, uses these texts to assess progress and plan subsequent therapy. We therefore treat these materials as texts for specific purposes, namely speech and language recovery.

Selecting thematically relevant texts for patients and having them expertly simplified is a task that exceeds the capacity of a practicing speech therapist in routine clinical settings (Hoover et al., 2023). While engaging reading programs show promise (Cocks et al., 2013), manual customization to adult interests and impairment-specific linguistic levels remains difficult to scale. We address this gap by presenting a sustainable computational pipeline for generating texts with specified topics, genre structures, and linguistic complexity, thereby creating an expandable repository of such materials.

2. Therapeutic Texts for Aphasia: Topic, Structure, and Complexity

To understand the specific features of texts used in aphasia therapy as a genre, we compiled a corpus of 127 texts. These were drawn from textbooks and methodological guides for speech-language therapists working with aphasia, as well as from texts employed by practicing therapists, selected by them to meet the needs of individual patients.

Further analysis revealed that texts designed for patient work predominantly feature child-oriented themes: parent–child relationships, children's games, and grandparents with grandchildren. However, in texts individually selected by speech therapists, the topics are already more closely

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 41–49

aligned with patients' age and life experience: cars, hobbies, gardening. In expert interviews, therapists noted that patients feel offended when offered child-oriented reading materials. We compiled a list of 12 topics designed to represent ecological domains of adult life: Domestic Activities, Modern Technology, Gardening, Hobbies, Cooking, Travel & Transport, Shopping & Finances, Health, Socializing, Pets, Work & Profession, and Daily Routine. This selection prioritizes subjects that facilitate functional communication and social engagement, purposely excluding child-oriented themes common in traditional therapy.

In terms of composition, texts for speech therapy are structured with remarkable clarity: exposition, climax, resolution.

Interestingly, there is a certain tonal movement within the text: it may begin neutrally, then acquire a slight negative shading, but the final part is always positive and contains a takeaway — a kind of lesson or moral arising from the situation described. The text prompts reflection, observation, or gentle irony — not didacticism, but a soft philosophical undertone. Meaning is conveyed through action or image rather than direct moralizing, allowing the reader to infer that "this is about us":

> **Example translated to English.** A gardener had two sons. The gardener owned a large vineyard. The sons did not want to work in the vineyard. One day, the father called his sons and said, "Children, go into the vineyard and look for something hidden there." The sons thought a treasure was buried in the vineyard. They went and began digging the ground. They dug up the entire vineyard, but found no treasure. However, the soil became loose and fertile. The vineyard produced abundant grapes. The sons sold the grapes and became rich.

Each text corresponds to one of the following genres: a real-life incident, a parable, a fairy tale, or a didactic story. There is a conflict or contrast (misunderstanding, change, confrontation). The hero loses something, discovers something, or undergoes a change (internal or external). The conflict is neither dramatic nor tragic. The number of characters in the texts is small — typically one or two main protagonists. To systematize them for generation purposes, we drew on the principles of the morphology of the folktale as developed by the Russian formalist Vladimir Propp (Propp, 1968). He argued that despite the immense variety of fairy-tale plots, seven generalized character types can be identified. An analysis of existing therapy texts allowed us to distinguish ten such character archetypes. They reflect the core gender and age oppositions that structure society, as well as two basic social roles from the "inner circle" (neighbor and friend): Old Man, Old Woman, Man, Woman, Boy, Girl, Friend, Neighbor, Young Man, Young Woman. In addition, the characters interact with some objects from the everyday environment. However, the object with which a character interacts is not merely named in the text — it is assigned a certain attribute. To systematize the attribute and its bearer for text generation, we employed a minimal predicative unit: the "noun + adjective" pair. The choice of the predicative word was likewise not arbitrary: according to the theory of argumentation within language developed by Jean-Claude Anscombre and Oswald Ducrot (Anscombre and Ducrot, 1983), every linguistic community possesses prototypical representations of the properties that objects and people can have. Thus, prototypical work is hard, and a prototypical relative is close. Drawing on our own introspection as native speakers and on association measures between collocates in the Russian National Corpus, we formulated eleven such "object — prototypical attribute" pairs: red ball, big book, old key, wooden box, heavy stone, important letter, old photo, yellow flower, blue cup, green apple, and new hat.

Regarding the linguistic complexity of texts for speech therapy, its parameters correspond to the severity of the patient's language impairment. However, in clinical practice, speech therapists adjust these parameters intuitively, based on their own expertise. In our previous work, we explicitly formulated three types of linguistic complexity requirements tailored to patients with mild, mild-to-moderate, and moderate impairment severity (Author and Author, 2024). In the following text, we will refer to these levels as Basic, Intermediate and Advanced, respectively. These requirements were developed through iterative collaboration with expert speech therapists and were clinically validated. When discussing the generation results, we will draw, in part, on these text requirements.

Thus, the analysis of texts already used in therapy, combined with input from speech therapists, allowed us to formulate the following aspects and characteristics of this text type, which will subsequently inform both the generation and validation of the output. Table 1 summarizes the target properties used as design constraints.

3. Methods

The input space is modelled as a semantic triplet (c, a, o), where c is a thematic category (12 types), a is a character archetype (10 types), and o is a key object (11 types). This combinatorial space ($1,320$ unique triplets) ensures thematic diversity

#	Aspect	Description
1	Structure	Clear three-part structure: exposition, climax, resolution
2	Character	10 archetypes reflecting core social roles
3	Key Object	11 object-attribute pairs for interaction
4	Conflict	Conflict or contrast (misunderstanding, change, confrontation)
5	Depth	Philosophical undertone / gentle reflection
6	Emotion	Socially approved emotions dominate
7	Genre	Real-life incident, parable, fairy tale, or didactic story
8	Language	3 regimes: Basic, Intermediate, Advanced
9	Topic	12 adult-oriented topics

Table 1: Target aspects and characteristics of therapeutic texts for aphasia.

while avoiding the infantilized narratives typical of accessible reading materials.

We operationalize clinical progression via three complexity regimes defined by constraint bundles. Basic texts are strictly limited to 5–7 simple past-tense SVO sentences with no dialogue. Intermediate narratives introduce limited subordination within a 2–3 paragraph structure, while Advanced texts permit complex syntax, descriptive modifiers, and abstract themes. This approach ensures clinical interpretability across diverse topics.

To validate these regimes, we test two hypotheses and one research question: (H1) complexity regimes form statistically distinct linguistic distributions; (H2) Basic texts align quantitatively with a clinical anchor; and (Q3) high-level narrative constraints require LLM-based assessment. Following (Alva-Manchego et al., 2020), we assess whether the regimes form statistically separable distributions across readability and syntactic indicators. To capture tonal alignment, we evaluated sentiment distributions using the `cointegrated/rubert-tiny-sentiment-balanced` model. Since surface statistics cannot capture narrative structure or tonal appropriateness, we complement them with a deterministic LLM-as-a-judge pipeline using binary criteria, interpreting results with awareness of potential model biases (Zheng et al., 2023).

4. Text Generation Pipeline

We generate texts via batched prompting using Gemini 1.5 Flash (Reid et al., 2024) through the Google Generative AI API (Figure 1).We selected this model because it matched the operational requirements of the generation stage:

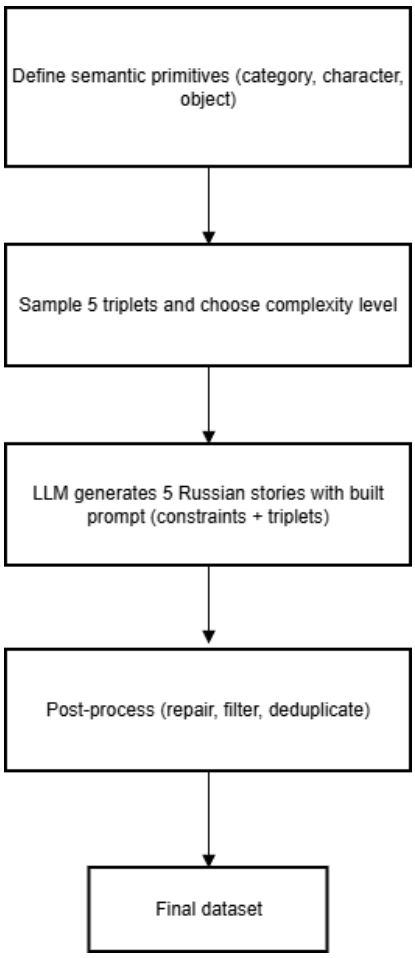

Figure 1: Synthetic dataset creation pipeline producing 1,296 unique stories.

stable schema-following behavior, efficient API-based batched inference, and sufficiently reliable instruction-following under Russian prompt conditions. Since Google does not publicly disclose the parameter count of Gemini 1.5 Flash, we avoid making capacity claims based on model size and motivate its use empirically through pilot-stage generation behavior instead. For each batch, we fix one complexity regime, sample five unused triplets, and prompt the model to produce one story per triplet under regime-specific constraints. We use deterministic or low-variance decoding and enforce a fixed output schema in the prompt.

Large-scale generation produces recurrent formatting failures. We observe CSV breaks caused by commas inside the generated text field; we repair these by merging spillover columns back into the text column until the expected schema is restored. We remove empty outputs and deduplicate near-identical stories.

After post-processing, the corpus contains 1,296 unique stories: 31.0% Basic, 33.7% Intermediate, and 35.3% Advanced.

Furthermore, the analysis revealed a high de-

gree of corpus balance with respect to content parameters as well. All 12 thematic categories are uniformly represented, each accounting for between 8.0% and 8.6% of the total number of texts. A similarly even distribution is observed for the 10 character archetypes and the 11 key objects. This balance ensures that when the corpus is used in therapeutic practice, the risk of systematic bias toward any particular theme, character, or object is minimized, thereby promoting greater variability and potential patient engagement.

5. Evaluation

We benchmark against a "Ground Truth" (GT) set of 10 manually authored therapy texts used in current clinical practice. These texts were selected by practicing therapists based on their everyday utility. They were not pre-selected to match any one generated regime; rather, we use them as a small clinical anchor representing materials currently employed in routine therapeutic work.

5.1. Classical Linguistic Metrics

We evaluated the full synthetic corpus and used the 10 GT texts to contextualize the regimes. We report readability (SMOG McLaughlin, 1969; Russian Flesch Reading Ease Oborneva, 2006), lexical diversity (MTLD McCarthy, 2005), and syntactic complexity (mean dependency distance, MDD Liu, 2008), alongside surface statistics.

As shown in Table 2, the results exhibit a monotonic progression from Basic to Advanced across all indicators, complete with standard deviations to capture within-level variability.

Regarding H2, the metrics reveal an important nuance: the Ground Truth texts do not uniformly align with the Basic generated level. While the GT texts closely match the Basic level in readability (Flesch RE: 68.0 vs. 65.0), they align much more closely with the *Intermediate* generated level in Word Count (92.8), SMOG (13.2), MTLD (123.8), and MDD (2.73). This indicates that the automatically generated "Basic" texts are structurally simpler (shorter and with lower dependency distance) than the therapist-authored clinical anchor. We therefore interpret the Basic regime as a lower-complexity option that may be more accessible for patients with more severe aphasia, rather than as a direct approximation of standard therapist-authored materials.

Additionally, sentiment analysis highlighted a distinct tonal shift. The human-authored GT texts feature notable negative dramatic tension (Negative Sentiment: 0.500 ± 0.33, Compound: -0.35), typical of fables that establish a problem before a resolution. The generated texts, however, lean heavily positive across all levels (Compound scores ranging from $+0.22$ to $+0.30$). To confirm that these observed differences reflect statistically distinct distributions rather than incidental variation, we applied the Kruskal–Wallis H test to core metrics (Kruskal and Wallis, 1952). As shown in Table 3, differences across regimes are highly significant for all tested indicators (all $p < 0.001$), supporting H1.

Metric	Basic	Interm.	Adv.	GT
Surface Indicators				
Word Count	23.0 ±3.1	59.2 ±9.2	111.0 ±31.5	92.8 ±35.6
ASL (words/sent)	3.4 ±0.5	10.3 ±2.4	13.2 ±2.7	9.2 ±2.8
Readability				
Flesch RE (RU)	65.0 ±24.7	52.6 ±12.8	49.3 ±11.0	68.0 ±9.2
SMOG	9.5 ±1.7	14.6 ±1.7	16.0 ±1.7	13.2 ±2.0
Lexical & Syntactic				
MTLD	23.4 ±17.3	137.1 ±66.7	154.2 ±42.1	123.8 ±56.3
MDD	1.54 ±0.15	2.83 ±0.26	3.02 ±0.24	2.73 ±0.31
Subclauses / sent	0.03 ±0.09	0.87 ±0.38	0.99 ±0.39	0.50 ±0.28
Morphological				
Past Tense Ratio	0.98 ±0.09	0.81 ±0.14	0.79 ±0.15	0.72 ±0.15
Errors / 100w	0.04 ±0.43	0.31 ±0.88	0.27 ±0.54	1.32 ±3.60

Table 2: Mean linguistic metrics (± SD) by complexity regime and ground-truth (GT) texts.

Metric	H	p
Average Sentence Length (ASL)	936.6	< 0.001
Mean Dependency Distance (MDD)	882.8	< 0.001
SMOG Readability Index	872.8	< 0.001

Table 3: Statistical validation of complexity stratification using the Kruskal–Wallis H test.

At higher levels, the model drifts toward a syntactically denser narrative style than the clinically curated GT: Intermediate and Advanced texts show substantially higher subordination rates (0.87 and 0.99 subclauses per sentence) than the GT (0.50). This suggests that, without explicit constraints, the model does not reliably preserve the unusually sparse syntax characteristic of the therapist-authored clinical anchor. This syntactic drift motivates complementary, structure-aware evaluation beyond surface statistics (Q3; see Section 5.2).

5.2. LLM-as-a-Judge

To address Q3, we implemented an automated evaluation pipeline using Gemma-3-4b-it to assess narrative coherence, tonal appropriateness, and fine-grained linguistic constraints (Gemma Team et al., 2024). We selected Gemma-3-4b-it as a compact open 4B instruction-tuned model for repeated rubric-based evaluation. Gemma 3 supports multilingual processing, including Russian, and provides a 128K context window, making it practical for consistent judge-style assessment across our sampled corpus. We treat these

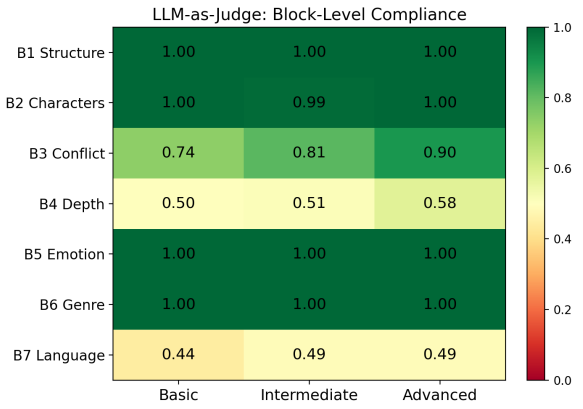

Figure 2: Heatmap of LLM-as-a-Judge compliance scores across seven conceptual blocks. (Note: 'short', 'medium', and 'long' in the figure correspond to the Basic, Intermediate, and Advanced regimes, respectively). Structural and emotional constraints show near-perfect adherence, while Language Constraints (B7) and Semantic Depth (B4) reveal limitations.

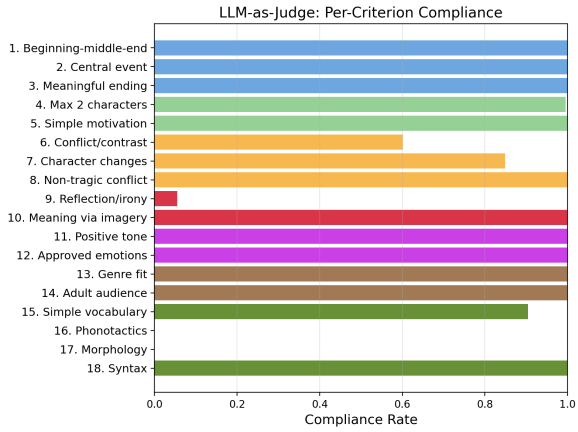

Figure 3: Detailed compliance rates by individual criterion (1–18). The chart illustrates the disparity between high-level narrative success and low-level linguistic control (e.g., Phonotactics).

judge scores as exploratory and complementary to classical metrics, not as a substitute for expert clinical assessment. We designed a rubric comprising 18 binary criteria organized into seven conceptual blocks (aligning directly with the aspects in Table 1). A random sample of 198 stories (stratified by complexity level, 66 per level) was evaluated using deterministic decoding to ensure reproducibility. The prompt explicitly instructed the LLM-as-a-Judge model to act as a harsh critic, verifying constraints specific to the Russian language (Zheng et al., 2024). Note that some criteria evaluated here (e.g., presence of reflection/irony, strict phonotactics) were not explicitly mandated in the generation prompt; this evaluation serves to identify natural gaps in LLM outputs to inform future neuro-symbolic interventions, rather than solely grading prompt adherence.

The overall compliance rate across the sampled corpus was 80.0%. At the block level (Figure 2), the model demonstrated near-perfect performance on high-level narrative instructions. Compliance scores for Narrative Structure (B1), Emotional Tone (B5), and Genre Appropriateness (B6) were nearly absolute across all levels. Notably, the Conflict & Dynamics block (B3) revealed a complexity-dependent progression: the model struggled to establish meaningful conflict within the tight constraints of Basic texts but successfully developed narrative tension when afforded greater length in Advanced stories.

However, the criteria-level analysis (specifically within B7: Language Constraints, see Figure 3) highlights specific limitations in controlling fine-grained linguistic features. The criterion prohibiting consonant clusters of three or more sounds (Phonotactics) yielded 0% compliance from the judge. This criterion is important because articulating such consonant clusters is problematic for patients with motor aphasia. To validate this result, we manually checked the corpus using a deterministic regex-based script. The manual analysis showed that actual compliance was 18.6% overall (Basic: 52.2%, Interm.: 6.6%, Adv.: 0.4%). Thus, although the generator does struggle with this constraint, the judge substantially underestimates compliance on this character-level task. We interpret this as evidence that LLM-based evaluators are unreliable for phonotactic verification of this kind and should be supplemented with deterministic symbolic checks.

Additionally, a discrepancy emerged in Morphology: while classical metrics confirmed high past-tense usage, the judge's binary scoring penalized even single deviations, resulting in a floor effect. Finally, the low score on Reflection/Irony suggests that while the model generates coherent therapeutic scenarios, it lacks the emergent semantic depth required to produce the philosophical subtext characteristic of human-authored fables without few-shot guidance.

Judge comparison against the clinical anchor. To better contextualize the judge-based scores, we also applied the same 18-criterion rubric to the 10 therapist-authored ground-truth (GT) texts currently used in clinical practice. As shown in Figure 4, the GT texts achieved an overall compliance rate of 83.3%, compared to 80.0% for the generated sample. This result is important for two reasons. First, it shows that the generated corpus remains close to the clinical anchor at the aggregate level under this rubric. Second, it reveals lim-

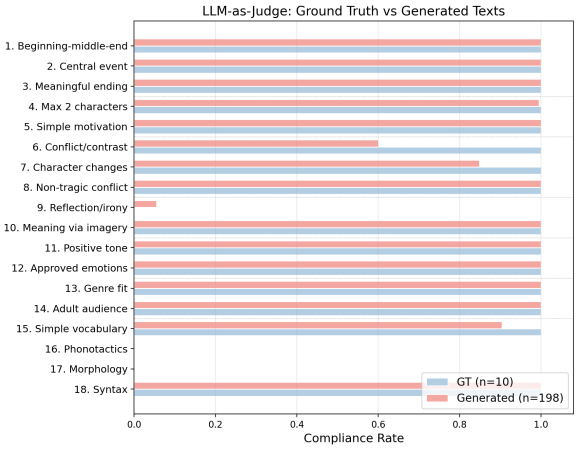

Figure 4: LLM-as-a-Judge comparison between therapist-authored ground-truth (GT) texts and generated texts. GT attains a slightly higher overall compliance rate (83.3% vs. 80.0%), while shared failures on criteria such as Phonotactics, Morphology, and Reflection/Irony expose limitations of the judge itself rather than only weaknesses of the generated corpus.

itations of the judge itself: for several criteria, the model penalizes properties that are also naturally present in therapist-authored materials.

In particular, the judge assigned 0% compliance to both GT and generated texts for Phonotactics and Morphology, despite the fact that such violations are also expected in authentic clinical texts and, in the case of phonotactics, were already shown to be underestimated by the LLM judge relative to manual verification. Likewise, the Reflection/Irony criterion remained near-zero even for GT texts, suggesting that this aspect is not robustly captured by the judge in Russian. Taken together, these results indicate that the LLM-as-a-judge protocol is informative for high-level structural and tonal screening, but should be interpreted cautiously for fine-grained linguistic constraints and abstract semantic properties.

6. Discussion

The results of this study highlight both the transformative potential and the inherent boundaries of using Large Language Models for clinical text generation. A primary finding is the distinct trade-off between therapeutic safety and narrative tension. While the generated texts partially align with the clinical anchor, particularly in readability at the Basic level, they remain structurally simpler than therapist-authored materials in length and syntactic complexity and exhibit a marked “positivity bias” introduced by the safety alignment (RLHF) of the underlying model (Ouyang et al., 2022). Ground-truth texts, often drawn from classical fables, rely on negative dramatic tension (compound sentiment -0.35) to drive the narrative. In contrast, the generated corpus leans heavily towards positive or neutral resolutions ($+0.22$ to $+0.30$). While this minimizes the risk of triggering emotional distress in patients with post-stroke depression, the lack of conflict in Basic texts (0.74 compliance) suggests that excessive safety constraints may inadvertently sterilize the narrative, potentially reducing patient engagement and motivation.

Furthermore, our analysis reveals a divergence in syntactic density at higher complexity levels. While the generated Basic texts mirrored (and even simplified) the syntactic structure of the clinical anchor, the Intermediate and Advanced levels exhibited subordination rates significantly exceeding the ground truth. This suggests that, without explicit constraints, the model tends to produce a denser narrative syntax than that observed in therapist-authored clinical materials. More broadly, complexity control in LLMs appears to function more as a ceiling than a floor: the model readily simplifies text when strongly constrained, but struggles to preserve the unusually sparse syntax characteristic of therapeutic materials at higher narrative lengths.

A further comparison with the therapist-authored GT texts reinforces this point. When evaluated under the same 18-criterion rubric, GT achieved only a slightly higher aggregate score than the generated corpus (83.3% vs. 80.0%). At the same time, several low-scoring criteria remained problematic for both text types, including Reflection/Irony, Phonotactics, and Morphology. This suggests that part of the observed error signal should be attributed not only to generation failures, but also to limited judge sensitivity for Russian fine-grained linguistic and semantic phenomena.

Finally, the error analysis underscores the ”tokenization wall” in prompt-based control. The failure to satisfy phonotactic constraints, combined with the LLM-judge’s inability to accurately assess them, confirms that token-based architectures cannot reliably filter sub-word phonetic features in-context (Sennrich et al., 2016). This necessitates a shift from purely generative pipelines to hybrid neuro-symbolic systems, where the LLM is responsible for semantic coherence and narrative structure, while deterministic post-processors enforce strict phonological and morphological constraints.

7. Conclusions

We present a scalable, controllable pipeline for generating therapeutic reading materials for Russian-speaking patients with aphasia. By formalizing the input space through semantic triplets

and operationalizing clinical progression into discrete complexity regimes, we demonstrated that Large Language Models can produce structurally sound and thematically diverse texts that adhere to strict linguistic constraints. Our evaluation confirms that the generated complexity levels are distinct, with the "Basic" level offering a substantially simpler structural profile, while the "Intermediate" level more closely matches the length and lexical diversity of the therapist-authored clinical anchor.

Future iterations of this work will focus on three key directions. First, we aim to integrate neuro-symbolic post-processing modules to strictly enforce phonotactic rules, addressing the current limitations in consonant cluster control. Second, we will expand the comparative evaluation on therapist-authored Ground Truth texts by increasing the size of the clinical anchor and by comparing judge-based scores with expert human ratings. Finally, and most critically, we will transition from computational evaluation to clinical validation. We plan to deploy the generated corpus in a pilot study with practicing speech therapists (logopedists) and patients to assess the material's impact on reading comprehension and engagement in a real-world rehabilitation context. This feedback loop will be essential for fine-tuning the generation parameters and moving from a synthetic dataset to a clinically deployable tool.

8. Limitations

Our findings should be interpreted in light of four limitations. (1) Small clinical reference set: the manually curated GT texts serve as a qualitative anchor rather than a representative baseline for Russian aphasia therapy materials; comparisons (e.g., sentiment and syntactic norms) are indicative. Scaling this reference requires digitizing a larger body of clinical manuals. (2) LLM-as-a-Judge bias: despite deterministic decoding and a rubric, judge-based scores may be affected by model bias (e.g., favoring texts aligned with its own distribution) and tokenization limits (as proven by our manual phonotactic verification). (3) Lack of expert validation: while we report a broad set of automated metrics, we have not yet conducted formal blinded expert assessment by practicing speech-language pathologists of the generated texts' overall therapeutic appropriateness and qualitative utility. (4) No ecological validation: we evaluate computational constraint satisfaction, not therapeutic effectiveness; readability metrics do not guarantee learnability or clinical utility, which also depend on presentation factors (layout, font) and patient-specific deficits.

9. Ethics Statement

This research involves the generation of synthetic texts intended for potential clinical use in aphasia rehabilitation. While these texts are designed to be safe, engaging, and devoid of sensitive or triggering topics, they are experimental and not a replacement for professional clinical judgment. Any deployment of these materials in actual therapy must be done under the strict supervision of qualified speech-language pathologists. Furthermore, the text generation process relied exclusively on generalized archetypes and predefined topics; no real patient data, clinical records, or personally identifiable information was used to train, prompt, or evaluate the models in this study.

10. Acknowledgements

This work was supported by the Basic Research Program at the National Research University Higher School of Economics.

11. Bibliographical References

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Jean-Claude Anscombre and Oswald Ducrot. 1983. *L'argumentation dans la langue*. Pierre Mardaga, Liege.

Author and Author. 2024. Anonymized. In *Anonymized*, pages 178–180. Anonymized for review.

Naomi Cocks, Martin Pritchard, Helen Cornish, Nicholas Johnson, and Madeline Cruice. 2013. A "novel" reading therapy programme for reading difficulties after a subarachnoid haemorrhage. *Aphasiology*, 27:509–531.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open models based on gemini research and technology.

Elizabeth Hoover, Ellen Bernstein-Ellis, and Debra Meyerson. 2023. Using bibliotherapy to rebuild identity for people with aphasia: A book club experience. *Journal of Communication Disorders*, 105:106363. Epub 2023 Jul 28. PMID: 37517172.

William H. Kruskal and W. Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Haitao Liu. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9(2):159–191.

A. R. Luria. 2008. *Vysshie korkovye funkcii cheloveka [Higher human cortical functions]*. Piter, St. Petersburg. (in Russian).

Philip M. McCarthy. 2005. An assessment of the range and limiting values of measures of lexical diversity. *Department of English, University of Memphis*. Dissertation.

G. Harry McLaughlin. 1969. Smog grading: A new readability formula. *Journal of Reading*, 12(8):639–646.

I. V. Oborneva. 2006. Modified flesch reading ease score for russian texts. *Automatized Analysis of Specialized Texts*, pages 23–28. (in Russian: Avtomatizirovannyy analiz spetsializirovannykh tekstov).

Long Ouyang, Jeffrey Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Vladimir Propp. 1968. *Morphology of the Folktale*. University of Texas Press, Austin, TX. Original work published 1928.

Machel Reid, Nikolay Savinov, Denis Teplyashin, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Janet Webster, Julie Morris, and David Howard. 2023. Reading comprehension in aphasia: the relationship between linguistic performance, personal perspective, and preferences. *Aphasiology*, 37(5):785–801.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A. Representative Generated Examples

To illustrate the controllable generation setup, we provide three representative examples, one for each complexity regime. Each example is annotated with its thematic category, main character, key object, and genre profile. For readability, this appendix includes English translations only; the original Russian versions are available in our repository.

The thematic categories used in the generation framework come from a fixed 12-topic inventory: Vehicles, Wild and Domestic Animals, Home and Garden, Games, Internet and Telecommunications, Arts and Entertainment, Computers and Electronics, AI, Beauty and Fitness, Travel, Sustainability, and Hobbies.

These examples are intended to demonstrate how semantic conditioning and regime-level complexity control interact in practice. They are illustrative and are not used as additional evaluation items.

A.1. Example A: Basic Regime

Metadata. Category: Home and Garden; Main Character: Boy; Key Object: Blue cup; Genre Profile: Real-life incident.

English translation

> When he came home, the father asked the children, “What did you do today, children?”
>
> “I washed the dishes,” Olya said proudly.
>
> “And I dried the dishes,” Katya said.
>
> The father turned to his son: “And you, Seryozha?”
>
> “And I picked up the broken pieces,” he replied.

Comment This example illustrates the lowest-complexity regime: a short event structure, an everyday setting, limited narrative development, and high interpretability.

A.2. Example B: Intermediate Regime

Metadata. Category: Home and Garden; Main Character: Girl; Key Object: Yellow flower; Genre Profile: Short narrative.

English translation

> A girl named Katya was helping her grandmother in the garden. Her grandmother was teaching her how to care for

plants. Katya especially loved the yellow flower growing by the fence. She watered it every day, carefully loosening the soil around it. One morning, Katya saw that the flower had bloomed. It was so bright and beautiful.

"Grandma, look how beautiful the flower is!" Katya exclaimed.

Her grandmother smiled: "That is the reward for your hard work, dear."

Comment This example illustrates the intermediate regime through a longer narrative arc, explicit causal progression, and a simple didactic ending while remaining lexically accessible.

A.3. Example C: Advanced Regime

Metadata. Category: Internet and Telecommunications; Main Character: Old Woman; Key Object: Wooden box; Genre Profile: Real-life incident.

English translation

Grandmother Agafya, a woman of about eighty, looked at her granddaughter with distrust.

"This wooden box," said the granddaughter, "is the Internet. With it, you will be able to talk to your son in America!"

The grandmother carefully took the box and touched its smooth surface. Her granddaughter patiently explained how to use the video call. At first, Agafya was frightened by every click, but when she finally saw her son's face on the screen, her eyes filled with tears of joy.

"Hello, Mother!" came her son's voice.

Smiling, the grandmother whispered, "Now this is a miracle!"

Comment This example illustrates the advanced regime through a longer narrative span, more elaborate scene-building, stronger emotional progression, and a clearer multi-step event structure.