



A Comparative Study of Synonym Generation in Russian with Instruction-Tuned Large Language Models in a Zero-Shot Setting

Maria Khokhlova
St Petersburg State University
m.khokhlova@spbu.ru

Can instruction-tuned LLMs reliably generate synonyms for Russian motion verbs —
and how does decoding temperature affect their output?

1. Research Questions

RQ1

Can instruction-tuned LLMs effectively generate synonyms for Russian verbs in a zero-shot setting?

RQ2

How do model type and decoding temperature affect the quality, diversity, and stability of generated synonyms?

2. Experimental Setup

18 Russian motion verbs
(from AG-80)

2 instruction-tuned LLMs
(Mistral / T-lite)

5 temperature settings
(0.0, 0.3, 0.5, 0.7, 1.0)

5 runs per configuration

8,217 generated items

3. Data and Models

18 Russian motion verbs
(from the Academic Grammar
of the Russian Language, AG-80):
*бежать, брести, везти, вести,
гнать, гнаться, ехать, идти,
катить, катиться, лезть,
летать, нести, нестись, плыть,
ползти, тащить, тащиться*

Models:

- Mistral-7B-Instruct-v0.3
(general-purpose multilingual)
- T-lite-instruct-0.1
(Russian-oriented)

Inference parameters:

- temperature: 0.0, 0.3, 0.5, 0.7, 1.0
- top_p: 0.7
- max_new_tokens: 50
- repetition_penalty: 1.05
- 5 runs per model × temperature × verb

4. Prompt (zero-shot)

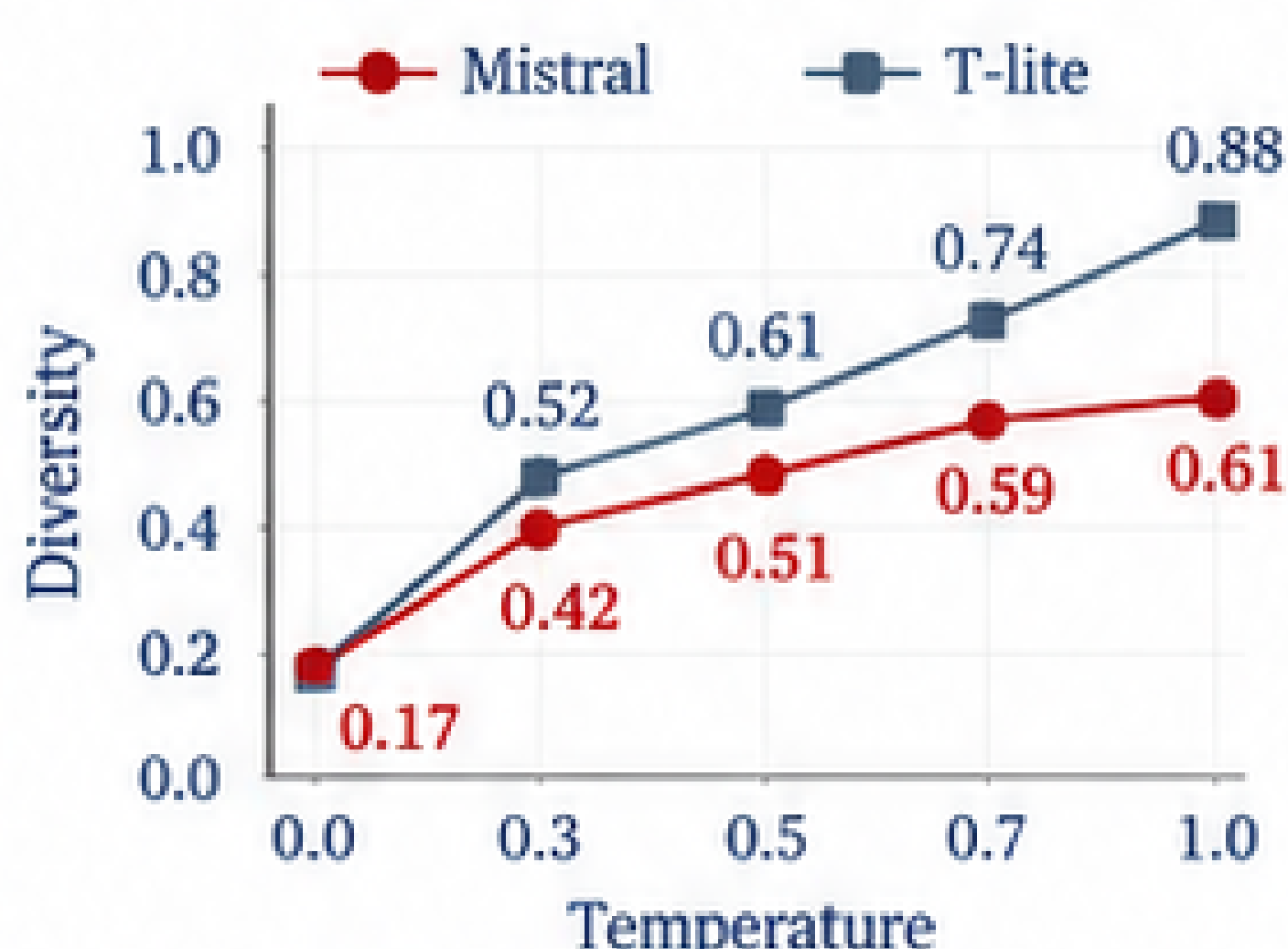
Generate 10 synonyms for each
verb. Synonyms must be verbs
in the infinitive, without
repetitions, ordered by
decreasing semantic similarity.

Required output format:

verb: synonym1, synonym2, ...

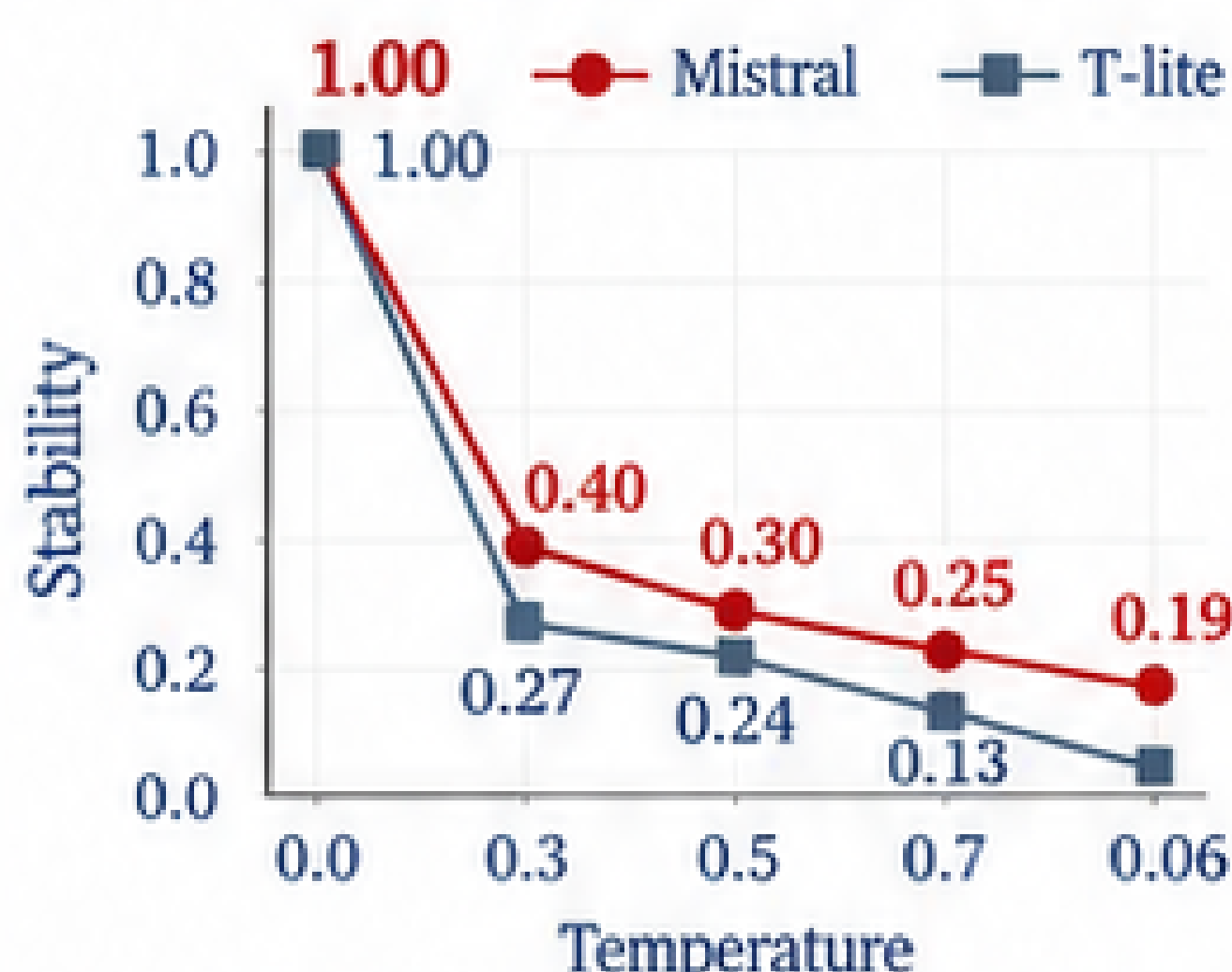
6. Results

6.1 Diversity increases with temperature



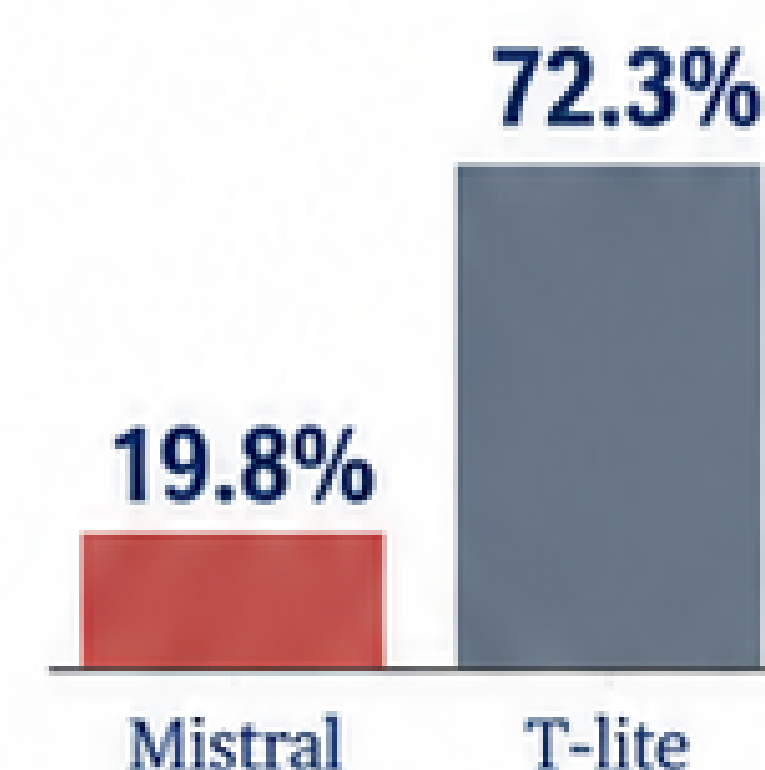
Both models become more diverse
as temperature increases, but the effect
is substantially stronger for T-lite.

6.2 Stability decreases with temperature



At t = 0.0, generation is highly stable;
increasing temperature progressively
reduces overlap between repeated runs.

6.3 Prompt adherence: exactly 10 synonyms



T-lite produced exactly
10 synonyms in 72.3%
of runs, compared to 19.8%
for Mistral.

5. Evaluation Metrics



Diversity

How varied are the outputs
across repeated runs?

$$D = \frac{|\bigcup W_i|}{\sum |W_i|}$$

Higher →
more unique
items



Stability

How similar are repeated
generations? Mean pairwise
Jaccard similarity.

$$S = \text{mean}_{i < j} \frac{|W_i \cap W_j|}{|W_i \cup W_j|}$$

Higher →
greater
overlap
across runs



Repeat frequency

How often does the same lexical
item reappear across runs?

Higher →
more recurrent
candidates

The metrics characterize generation behaviour
rather than semantic correctness; all generated
items were included in the calculations.

↑ TEMPERATURE → ↑ DIVERSITY but ↓ STABILITY

7. Qualitative Findings

T-lite: semantically coherent lexical candidates

бежать 'to run'

→ бегать, мчаться, нестись, убежать, спешить

брести 'to trudge'

→ плестись, ковылять, тащиться, волочиться

T-lite captures fine-grained semantic distinctions
related to the manner of motion.

Mistral: frequent deviations from the lexical task

Multiword expressions

бегать быстро 'to run fast'
катиться по кругу 'to roll in a circle'

Meaning descriptions

ползти → 'to move using arms and legs'

Malformed forms

агать · ваться · довать · вать · ься

Wrong language / POS

English candidates; noun погонщик
'herder/driver' instead of a verb.

8. Limitations / Error Analysis (T-lite)

Polysemy

тащиться 'to trudge' → врубаться
'to get into / understand (slang)'

The model activates a colloquial secondary
meaning rather than the target motion sense.

Semantic error

везти 'to transport by vehicle' →
говорить 'to speak'

Text continuation

вести → вести к, вести через
взбираться → взбираться на

T-lite occasionally generates semantically inappropriate
items or continues the input prompt, reflecting sensitivity
to highly frequent colloquial senses.

9. Conclusion

- Instruction-tuned LLMs can capture synonymic relations between Russian motion verbs in a zero-shot setting.
- Decoding temperature strongly affects generation behaviour: higher temperatures increase diversity but reduce stability.
- Intermediate temperature values provide a balance between diversity and stability.
- T-lite produces fewer irrelevant outputs and follows lexical constraints more successfully than multilingual Mistral.

Future work

Comparison with synonymic relations
documented in lexicographic resources
and evaluation of additional Russian-
oriented instruction-tuned LLMs.

Acknowledgment

Russian Science Foundation,
project No. 22-18-00189-P.