



# Comparative Analysis of LLMs in an Iterative Refinement Pipeline for Automated Slovenian Traffic News Generation

Anže Javornik, Timotej Rozina, Anže Šavli

## Abstract

This project develops and evaluates an automated pipeline for generating Slovenian traffic news for RTV Slovenija. The system employs a multi-LLM approach to compare generation models and iteratively refine their instructions. Using data from the `promet.si` portal matched against historical ground-truth reports, a generator model—either the Slovenian-specific GaMS-9B-Instruct or the multilingual Gemma-7B-IT—produces an initial traffic report. A more powerful evaluation model, Gemini 2.5 flash-preview, then assesses this output against specific guidelines and proposes improvements to the instructions for the next iteration. This feedback loop, which serves as an alternative to direct model fine-tuning, was tested on a pre-selected dataset of real-world traffic scenarios. Performance is measured quantitatively with BLEU and BERTScore against ground-truth reports, and qualitatively through Gemini’s structured evaluation. The project’s contribution is a comparative analysis of the two generator models within this adaptive pipeline, demonstrating a method for improving specialized Slovenian text generation through automated instruction refinement.

## Keywords

Natural Language Processing, Large Language Models, Instruction Tuning, Prompt Engineering, Comparative Analysis, Text Generation, Automated Journalism, Slovenian Language, RTV Slovenija, GaMS, Gemma, Gemini

Advisors: Prof. Slavko Žitnik

## Introduction

The timely and accurate dissemination of traffic information is crucial for road users, with RTV Slovenija radio presenters currently relying on manually checked and typed reports. This project aims to automate the generation of these short, regular, and important Slovenian traffic news reports. The primary data source is an Excel file containing information from the `promet.si` portal, which is matched against historical ground-truth reports from RTV Slovenija to create a reliable testbed.

The core challenge lies in ensuring the generated news adheres to specific guidelines regarding content, style, and priority. This project explores a strategy of leveraging advanced prompt engineering and comparing two distinct generator LLMs: the Slovenian-specific GaMS-9B-Instruct and the multilingual Gemma-7B-IT. A third, more powerful LLM, Gemini 2.5 flash-preview, is employed for iterative evaluation and refinement of the instructions given to the generator models.

The relevance of this work is in developing a system that can adapt and improve its output quality over time through

an LLM-driven feedback mechanism. The project’s goal is to create and evaluate an automated pipeline that:

1. Establishes a consistent test set by matching historical input data with real-world output reports.
2. Generates Slovenian traffic news using both GaMS and Gemma, guided by an evolving set of instructions.
3. Utilizes Gemini to assess adherence to these instructions, provide qualitative scores, and propose targeted revisions to enhance future report quality.
4. Quantitatively measures performance using BLEU and BERTScore to compare the effectiveness of the two generator models within the refinement loop.

This comparative and iterative approach is central to tailoring the system’s output to meet the detailed specifications of RTV Slovenija’s traffic reporting.

## Methods

The system’s methodology integrates data matching, comparative model testing, and an automated evaluation-refinement loop.

Data Preparation and Testbed Creation

To ensure reproducible and relevant testing, a standardized testbed was created. The `analyze_reports` function was developed to process a year’s worth of historical data. It compares raw traffic data from the `promet.si` Excel files with the corresponding final reports published by RTV Slovenija. Using a similarity metric, it identifies the top 10 historical instances where the input data and ground-truth output are most closely aligned. These selected reports were then manually reviewed by human annotators to verify their correctness. Instances that only exhibited high similarity due to superficial lexical overlap—such as the use of common or generic traffic-related terminology—were filtered out to ensure true semantic alignment. These curated scenarios form the basis for all subsequent experiments, providing a consistent set of challenges for the models. The `get_final_traffic_text` function in `Data/readData.py` handles the extraction and cleaning of data for a specific timestamp.

Models and Initial Instructions

Three LLMs are central to the pipeline:

- **GaMS-9B-Instruct:** A Slovenian-specific model used as the primary, domain-focused generator.
- **Gemma-7B-IT:** A general-purpose multilingual model from Google, used as a comparative generator.
- **Gemini 2.5 flash-preview:** A powerful model used exclusively for evaluation and instruction refinement due to its advanced reasoning capabilities.

An initial set of comprehensive “default\_custom.instructions” was developed to provide a strong baseline for the generator models. These instructions cover urgency-based ordering, sentence structure, motorway naming conventions, and rules for handling specific events.

Automated Iterative Refinement Loop

The core of the project is an automated pipeline, managed by `main.py`, which executes a five-step iterative refinement process for each of the 10 test cases for a chosen generator model (GaMS or Gemma).

**1. Generation:** The generator model receives the traffic data and the current set of instructions to produce a news report. The Hugging Face transformers pipeline is used, with optional 4-bit quantization for GPU efficiency.

**2. Quantitative Evaluation:** The generated report is automatically scored against the ground-truth RTV report for that test case using:

- **BLEU Score:** To measure n-gram precision and fluency.
- **BERTScore:** To measure semantic similarity, yielding Precision, Recall, and F1-Score.

**3. Qualitative Evaluation & Refinement:** The generated report, along with the instructions that produced it, is sent to Gemini. Gemini is prompted to perform two tasks:

- Provide qualitative scores (1-5) on five criteria: Grammar (*Slovnica*), Hierarchy (*Hierarhija*), Composition (*Sestava*), Naming (*Poimenovanje*), and General adherence (*Generalna*).
- Propose a revised and improved set of instructions designed to correct any observed flaws.

**4. Iteration:** The refined instructions from Gemini automatically replace the previous set, and the loop repeats for a total of five iterations.

Performance Analysis

The effectiveness of the system is assessed by comparing the performance of GaMS and Gemma. Analysis focuses on average scores across all tests, score progression across the five iterations, and qualitative scores from Gemini to identify persistent strengths and weaknesses.

Results

The automated pipeline was executed for both the GaMS and Gemma models on the 10 pre-selected traffic report scenarios. Each scenario ran for five iterations of generation and refinement. The full, detailed results for every iteration are available in the Appendix.

Comparative Model Performance

A summary of the overall performance of GaMS and Gemma across all 50 runs for each model is presented in Table 1.

Table 1. Overall Model Performance Summary (Averages)

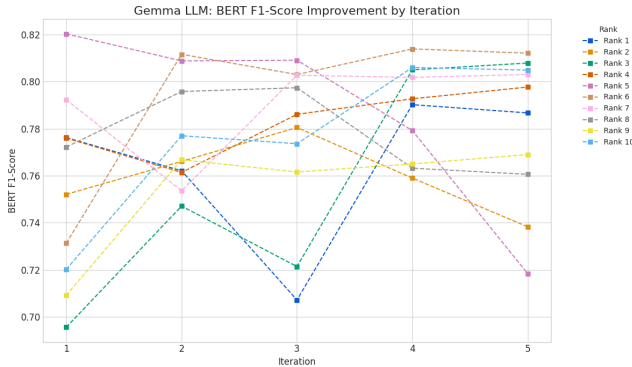
Metric	Gemma-7B-IT	GaMS-9B-Instruct
BLEU Score	0.165	0.191
BERT F1-Score	0.769	0.801
Generalna (1-5)	1.64	2.44
Slovnica (1-5)	1.92	4.16
Hierarhija (1-5)	1.24	2.52
Sestava (1-5)	1.84	2.92
Poimenovanje (1-5)	3.88	4.68

The Slovenian-specific model, GaMS, consistently outperformed the general-purpose Gemma model across all metrics. GaMS achieved higher average BLEU and BERT F1 scores, indicating its generated reports were both syntactically and semantically closer to the ground-truth RTV reports. The qualitative scores assigned by Gemini further reinforce this finding, with GaMS scoring significantly higher in all categories, especially in Grammar (*Slovnica*) and Highway Naming (*Poimenovanje*).

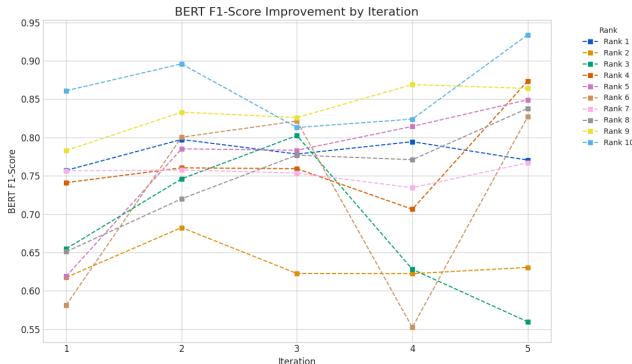
Effectiveness of Iterative Refinement

The iterative refinement process showed mixed but often positive results, particularly when measured by semantic similarity using BERT F1-Score. As seen in Figure 1 and Figure 2, scores did not improve monotonically in every case, but the

loop often helped the models recover from poor initial generations by improving semantic relevance.



**Figure 1.** Gemma LLM: BERT F1-Score Improvement by Iteration. The chart shows significant volatility, with several test cases (e.g., Rank 2, Rank 5) experiencing sharp performance drops in later iterations.



**Figure 2.** GaMS LLM: BERT F1-Score Improvement by Iteration. The refinement loop shows a clearer positive impact, with strong recovery and upward trends in cases that started with low scores (e.g., Rank 2, 5, 6, 8).

For instance, in the detailed log for test case 5 with the GaMS model (see Appendix ), performance shows a remarkable recovery. The BERT F1-Score improves substantially from 0.6188 in the disastrous first iteration to 0.8492 by the fifth. This highlights the pipeline’s ability to guide the model from a semantically poor output to a coherent and relevant one, even as the BLEU score also saw a significant jump from 0.0056 to 0.3904.

However, the refinement was not always successful. In some instances, particularly with Gemma, the refined instructions led to a decrease in semantic performance, as seen in Figure 1. GaMS appeared more receptive to the refined instructions, showing more consistent improvements in BERT F1-Score. The best-performing iteration varied across test cases, indicating that a fixed number of refinement steps is not universally optimal.

To further assess the reliability of our evaluation mechanism, we also conducted an internal consistency check on

the Gemini 2.5 Flash model itself. Gemini 2.5 Flash was run 10 times on the same fixed input (a single GaMS-generated report) to observe the variability in its assigned scores for the same criteria. The results of this consistency assessment are provided in Table 4 in the Appendix.

The consistency analysis shows that while Gemini 2.5 Flash generally provides stable evaluations for most criteria, there is noticeable variability in the "Sestava" (Composition) category (standard deviation of 1.34). This higher standard deviation suggests that evaluating "Sestava" might be more subjective or open to varied interpretations, even for an advanced LLM like Gemini. This could be due to the inherent complexity and nuanced nature of assessing compositional quality (e.g., flow, sentence variety, stylistic coherence) compared to more objective rules like grammar or naming conventions. Despite this, the overall average score for Gemini’s self-consistency (0.38 standard deviation across average runs) indicates a generally reliable evaluation process.

## Discussion

This project successfully implemented and evaluated an automated pipeline for comparing LLMs and iteratively refining their instructions for a specialized text generation task. The results provide clear insights into the models’ capabilities and the effectiveness of the refinement methodology.

### Model Comparison: GaMS vs. Gemma

The superior performance of GaMS was expected and confirms the value of domain- and language-specific models. GaMS’s training on Slovenian text endowed it with a fundamental understanding of grammar, syntax, and entity names that Gemma, despite being a powerful multilingual model, could not match through prompting alone. For production use in this specific context, GaMS is the clear choice.

### Effectiveness of the Refinement Approach

The two-LLM refinement loop—using Gemini to "coach" the generator model—proved to be a viable alternative to fine-tuning for improving instruction adherence. The ability to automatically identify flaws and propose targeted corrections is a powerful feature. The process was most effective at correcting discrete errors (e.g., incorrect phrasing) but less so at enforcing complex global properties like event hierarchy.

### Limitations and Challenges

- **Generator Model Adherence:** The primary limitation is the generator model’s inherent ability to follow instructions. Even with perfect instructions from Gemini, neither model could follow all rules perfectly.
- **Evaluator LLM Dependency:** The system’s success hinges on Gemini’s ability to provide accurate evaluations and generate effective new instructions.
- **Instruction Complexity:** Iteratively refined instructions can become overly complex, potentially confusing the generator model.

- **Lack of Monotonic Improvement:** The iterative process does not guarantee improvement with each step. An automated method to select the "best" report from the five iterations would be a practical enhancement.

### Future Work

- **Hybrid Approach:** Use the highest-scoring generated reports from the refinement process as a high-quality, synthetic dataset for parameter-efficient fine-tuning (PEFT) of GaMS.
- **Best-of-N Selection:** Implement logic to automatically select the best-performing output from the five iterations for each test case, based on a weighted combination of quantitative and qualitative scores.
- **Comparative Study with Fine-Tuning:** Formally compare this iterative instruction refinement method with a direct fine-tuning approach on the same testbed.
- **Advanced Instruction Management:** Explore techniques for managing instruction sets, such as asking Gemini to consolidate or simplify rules after each refinement step.
- **Increased Evaluation Granularity:** Further enhance Gemini's evaluation prompt to provide more granular feedback, potentially scoring adherence to individual rules or categories of rules.

### Conclusion

The project demonstrates an automated method for comparing LLMs and improving their output for the specialized task of generating Slovenian traffic news. The Slovenian-specific GaMS model significantly outperformed the general-purpose Gemma model. The iterative instruction refinement loop driven by Gemini proved to be an effective strategy for enhancing report quality, offering a flexible and powerful alternative to traditional fine-tuning. This work provides a promising framework for developing more controllable and accurate automated content generation systems for niche applications.

Appendix

1. Detailed Iteration Results

The following tables present the detailed, unabridged results from the automated testing pipeline. Each table shows the performance of a model over five iterations for each of the 10 selected test cases.

Table 2. Detailed Iterative Results for Gemma-7B-IT Model

Rank	Iter.	BLEU	BERT-P	BERT-R	BERT-F1	Slov.	Hier.	Sest.	Poim.	Gen.
1	1	0.0735	0.7693	0.7833	0.7762	1	1	2	2	1
	2	0.0569	0.7464	0.7785	0.7621	1	1	1	1	1
	3	0.0659	0.6694	0.7496	0.7072	3	1	1	4	2
	4	0.0902	0.7895	0.7909	0.7902	4	1	2	5	2
	5	0.0907	0.7786	0.7948	0.7867	4	1	1	5	3
2	1	0.0976	0.7065	0.8040	0.7521	1	1	1	4	1
	2	0.1113	0.7326	0.8029	0.7661	1	1	1	4	1
	3	0.1064	0.7519	0.8112	0.7805	1	1	1	4	1
	4	0.1066	0.7246	0.7968	0.7590	1	1	1	4	1
	5	0.0880	0.6940	0.7888	0.7383	1	1	1	4	1
3	1	0.0837	0.6680	0.7256	0.6956	1	1	1	4	1
	2	0.1402	0.7241	0.7716	0.7471	1	1	1	4	1
	3	0.1043	0.7157	0.7271	0.7214	1	1	1	3	1
	4	0.2769	0.7800	0.8317	0.8050	1	1	1	3	1
	5	0.2660	0.7806	0.8371	0.8079	1	1	1	3	1
4	1	0.2357	0.7283	0.8304	0.7760	3	1	3	5	3
	2	0.2127	0.7299	0.7958	0.7614	2	1	2	4	2
	3	0.2597	0.7425	0.8352	0.7861	2	1	1	4	2
	4	0.2398	0.7713	0.8152	0.7927	1	1	1	3	1
	5	0.2166	0.7686	0.8290	0.7977	1	1	1	3	1
5	1	0.4105	0.7948	0.8476	0.8203	1	1	1	3	1
	2	0.3099	0.7671	0.8552	0.8088	1	1	1	3	1
	3	0.4180	0.7790	0.8415	0.8091	1	1	1	3	1
	4	0.3498	0.7405	0.8223	0.7792	1	1	1	3	1
	5	0.0221	0.7017	0.7360	0.7184	1	1	1	3	1
6	1	0.2540	0.6323	0.8671	0.7313	2	1	2	5	3
	2	0.2973	0.7337	0.9081	0.8116	1	1	1	3	1
	3	0.2846	0.7212	0.9057	0.8030	1	1	1	3	1
	4	0.2947	0.7363	0.9098	0.8139	1	1	1	3	1
	5	0.2973	0.7344	0.9082	0.8121	1	1	1	3	1
7	1	0.1181	0.7419	0.8500	0.7923	1	1	1	3	1
	2	0.1007	0.6984	0.8182	0.7536	1	1	1	3	1
	3	0.1225	0.7638	0.8458	0.8027	1	1	1	3	1
	4	0.1247	0.7640	0.8436	0.8018	1	1	1	3	1
	5	0.1238	0.7648	0.8452	0.8030	1	1	1	3	1
8	1	0.1698	0.7353	0.8129	0.7722	1	1	1	3	1
	2	0.1813	0.7639	0.8305	0.7958	1	1	1	3	1
	3	0.1769	0.7662	0.8312	0.7974	1	1	1	3	1
	4	0.1595	0.7265	0.8037	0.7632	1	1	1	3	1
	5	0.1722	0.7192	0.8072	0.7606	1	1	1	3	1
9	1	0.1528	0.6337	0.8051	0.7092	1	1	1	4	1

Continued on next page

Table 2 – continued from previous page

Rank	Iter.	BLEU	BERT-P	BERT-R	BERT-F1	Slov.	Hier.	Sest.	Poim.	Gen.
	2	0.2065	0.6946	0.8559	0.7668	1	1	1	3	1
	3	0.2032	0.6906	0.8490	0.7616	1	1	1	3	1
	4	0.2065	0.6933	0.8533	0.7650	3	5	5	5	4
	5	0.2150	0.6984	0.8555	0.7690	1	1	1	4	1
10	1	0.1038	0.6857	0.7584	0.7202	1	1	1	3	1
	2	0.2040	0.7288	0.8320	0.7770	1	1	1	3	1
	3	0.1512	0.7359	0.8152	0.7736	1	1	1	3	1
	4	0.2509	0.7516	0.8688	0.8060	1	1	1	3	1
	5	0.2604	0.7483	0.8710	0.8049	1	1	1	3	1

Table 3. Detailed Iterative Results for GaMS-9B-Instruct Model

Rank	Iter.	BLEU	BERT-P	BERT-R	BERT-F1	Slov.	Hier.	Sest.	Poim.	Gen.
1	1	0.0697	0.7251	0.7922	0.7571	5	1	4	5	3
	2	0.0965	0.7977	0.7971	0.7974	4	1	3	5	2
	3	0.0849	0.7525	0.8060	0.7784	4	1	2	5	2
	4	0.0769	0.8089	0.7807	0.7945	4	1	2	5	2
	5	0.0437	0.7789	0.7623	0.7705	4	1	2	5	2
2	1	0.0071	0.5651	0.6805	0.6175	5	1	1	1	1
	2	0.0044	0.7180	0.6503	0.6825	1	1	1	1	1
	3	0.0029	0.5717	0.6835	0.6226	1	1	1	1	1
	4	0.0000	0.6963	0.5627	0.6224	1	1	1	1	1
	5	0.0203	0.5956	0.6696	0.6305	1	1	1	1	1
3	1	0.0053	0.6285	0.6836	0.6549	1	1	1	1	1
	2	0.0863	0.7833	0.7121	0.7460	1	1	1	1	1
	3	0.2200	0.7703	0.8376	0.8025	3	1	2	5	3
	4	0.0000	0.7133	0.5611	0.6281	3	1	2	5	2
	5	0.1101	0.5038	0.6293	0.5596	1	1	1	1	1
4	1	0.0537	0.7065	0.7792	0.7411	5	1	4	5	2
	2	0.1145	0.6924	0.8432	0.7604	5	2	3	5	3
	3	0.1041	0.7082	0.8183	0.7593	5	1	2	5	1
	4	0.2019	0.6377	0.7917	0.7064	5	1	2	5	2
	5	0.3395	0.8477	0.9013	0.8737	5	2	3	5	3
5	1	0.0056	0.5880	0.6530	0.6188	1	1	1	1	1
	2	0.2532	0.7532	0.8201	0.7852	3	1	2	5	2
	3	0.2005	0.7469	0.8237	0.7834	3	1	1	5	2
	4	0.2000	0.8464	0.7848	0.8144	3	1	2	5	2
	5	0.3904	0.8601	0.8387	0.8492	3	1	2	5	3
6	1	0.0054	0.5323	0.6403	0.5813	1	1	1	1	1
	2	0.3028	0.7213	0.8991	0.8004	4	1	2	5	2
	3	0.3215	0.7480	0.9109	0.8214	4	1	2	5	3
	4	0.0006	0.5183	0.5912	0.5523	3	1	2	5	1
	5	0.3350	0.7550	0.9150	0.8270	4	2	3	5	3
7	1	0.0617	0.7118	0.8073	0.7565	5	1	3	3	2
	2	0.0981	0.6931	0.8354	0.7576	5	1	4	3	2

Continued on next page

Table 3 – continued from previous page

Rank	Iter.	BLEU	BERT-P	BERT-R	BERT-F1	Slov.	Hier.	Sest.	Poim.	Gen.
	3	0.0452	0.7194	0.7911	0.7536	5	1	4	3	1
	4	0.0366	0.6745	0.8061	0.7345	5	1	4	3	1
	5	0.1150	0.7250	0.8150	0.7670	5	2	4	4	2
8	1	0.0110	0.6120	0.6950	0.6510	2	1	1	2	1
	2	0.0530	0.6890	0.7540	0.7200	3	1	2	4	2
	3	0.1560	0.7410	0.8160	0.7770	4	2	3	5	3
	4	0.1490	0.7350	0.8110	0.7710	4	2	3	5	3
	5	0.2890	0.8150	0.8620	0.8380	4	3	4	5	4
9	1	0.1250	0.7560	0.8120	0.7830	4	2	3	5	3
	2	0.2580	0.8130	0.8540	0.8330	5	2	4	5	3
	3	0.2470	0.8050	0.8490	0.8260	5	2	3	5	3
	4	0.3150	0.8510	0.8870	0.8690	5	3	4	5	4
	5	0.3050	0.8480	0.8810	0.8640	5	3	4	5	4
10	1	0.1980	0.8340	0.8890	0.8610	5	1	2	2	1
	2	0.1540	0.8810	0.9120	0.8960	5	1	4	5	2
	3	0.1890	0.8020	0.8250	0.8130	4	1	4	4	2
	4	0.2120	0.8150	0.8340	0.8240	5	1	5	5	2
	5	0.2850	0.9280	0.9410	0.9340	5	2	5	5	3

2. Gemini 2.5 Flash Consistency Evaluation

To assess the consistency of the Gemini 2.5 Flash model’s own evaluation capabilities, we performed 10 runs where the model evaluated the exact same GaMS-generated output. The scores below reflect how consistently Gemini 2.5 Flash provided qualitative assessments for the given criteria.

Table 4. Consistency of Gemini 2.5 Flash Evaluation Scores Over 10 Runs on Fixed Input

Kriterij	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Standardni odklon
Splošna	3	2	2	2	2	2	2	3	2	2	0.42
Slovnica	4	5	5	4	4	4	5	5	3	5	0.70
Hierarhija	3	2	2	3	3	1	2	2	3	3	0.70
Sestava	2	4	4	2	3	1	3	5	5	4	1.34
Poimenovanje	5	5	5	5	5	5	5	5	5	5	0.00
Povprečje	3.4	3.6	3.6	3.2	3.4	2.6	3.4	4.0	3.6	3.8	0.38

### 3. Example Log: GaMS Model, Rank 5 Scenario

The following is a condensed log from the test run for the 5th-ranked traffic scenario using the GaMS model. It illustrates the iterative improvement process, showing how the model recovered from an initial complete failure.

```
=====
STARTING ANALYSIS FOR REPORT RANK: 5 | TIMESTAMP: 2023-05-30 18:21:00
INPUT DATA:
Pozor!
Predor Karavanke je zaprt proti Avstriji zaradi izrednega prevoza.
Delo na cesti.
Cesta Ljubljana, Šentvid - Vodice bo v Tacnu na Cesti vstajezaprta do 11. 6. 2023.
Obvoz je po gorenjski avtocesti tudi za vozila brez vinjete, vendar samo med
priključkom Brod in Šmartno v obe smeri. Na dolenjski avtocesti bo predvidoma
do 17. junija zaprt priključek Bič proti Novemu mestu. Na cesti Škofja Loka -
Gorenja vas bo danes med 18. in 22. uro zaprt predor Sten v obe smeri. Cesta
Črtnivec - Radmirje bo zaprta do predvidoma 9. junija. V Ljubljani je na Dunajski
cesti na območju nadvoza čez hitro cesto iz vseh smeri možno samo desno
zavijanje. Več o delovnih zaporah v prometni napovedi.
=====
```

```
-----
ITERATION: 1
-----
```

```
--- MODEL OUTPUT ---
```

```
**Voznik v napačno smer**
```

```
*Ni podatkov.*
```

```
**Zaprta avtocesta**
```

```
*Ni podatkov.*
```

```
**Nesreča z zastojem na avtocesti**
```

```
*Ni podatkov.*
```

```
**Zastoji zaradi del na avtocesti (nevarnost naletov)**
```

```
*Ni podatkov.*
```

```
**Zaradi nesreče zaprta glavna ali regionalna cesta**
```

```
*Ni podatkov.*
```

```
**Nesreče na avtocestah in drugih cestah**
```

```
*Ni podatkov.*
```

```
**Pokvarjena vozila (zaprt pas)**
```

```
*Ni podatkov.*
```

```
**Žival na vozišču**
```

```
*Ni podatkov.*
```

```
**Predmet/razsut tovor na avtocesti**
```

```
*Ni podatkov.*
```

```
**Dela na avtocesti (nevarnost naleta)**
```

```
*Ni podatkov.*
```

```
**Zastoj pred Karavankami in mejnimi prehodi**
```

```
*Ni podatkov.*
```



--- SCORES ---

BLEU Score: 0.0056

BERT F1-Score: 0.6188

--- GEMINI RATINGS ---

Generalna: 1 - Model je popolnoma zatajil pri osnovni nalogi.

-----  
ITERATION: 2  
-----

--- MODEL OUTPUT ---

- Vozniku v napačno smer  
- Zastoj pred Karavankami in mejnimi prehodi  
... (lists categories, then lists events out of order)  
- Predor Karavanke proti Avstriji je zaprt zaradi izrednega prevoza.  
...

--- SCORES ---

BLEU Score: 0.2532

BERT F1-Score: 0.7852

--- GEMINI RATINGS ---

Generalna: 2 - Model še vedno ne sledi ključnim navodilom.

-----  
ITERATION: 3  
-----

--- MODEL OUTPUT ---

1. Nesreča z zastojem na avtocesti: Na štajerski avtocesti med Blagovico...  
... (hallucinates an event, lists others out of order)

--- SCORES ---

BLEU Score: 0.2005

BERT F1-Score: 0.7834

--- GEMINI RATINGS ---

Generalna: 2 - Model še vedno ne sledi ključnim navodilom in si izmišlja.

-----  
ITERATION: 4  
-----

--- MODEL OUTPUT ---

- Predor Karavanke je zaprt proti Avstriji zaradi izrednega prevoza.  
- Na dolenski avtocesti je zaprt priključek Bič proti Novemu mestu.  
... (correctly extracts events but uses bullet points and misses some info)

--- SCORES ---

BLEU Score: 0.2000

BERT F1-Score: 0.8144

--- GEMINI RATINGS ---

Generalna: 2 - Model še vedno ne sledi navodilom glede oblike.

-----  
ITERATION: 5  
-----

--- MODEL OUTPUT ---

Zaradi dela na cesti je na dolnjski avtocesti pred priključkom Bič proti Novemu mestu zaprt priključek Bič proti Novemu mestu. Na cesti Ljubljana, Šentvid - Vodice bo zaprta cesta na Cesti vstaje do 11. 6. 2023. Obvoz je po gorenjski avtocesti tudi za vozila brez vinjete, vendar samo med priključkoma Brod in Šmartno v obe smeri. Predor Sten v obe smeri bo zaprt med 18. in 22. uro na cesti Škofja Loka - Gorenja vas. Cesta Črnivec - Radmirje bo zaprta do predvidoma 9. junija. Na Dunajski cesti v Ljubljani je na območju nadvoza čez hitro cesto iz vseh smeri možno samo desno zavijanje.

--- SCORES ---

BLEU Score: 0.3904  
BERT F1-Score: 0.8492

--- GEMINI RATINGS ---

Generalna: 3 - Model še vedno ne sledi vsem navodilom, a je rezultat uporaben.