



# Automatic generation of Slovenian traffic news for RTV Slovenija

Žan Pušenjak, Matic Zadobovšek, Janez Tomšič

## Abstract

In this project we aimed to write traffic reports with the help of generative models. To achieve the best results we explored prompting, model instructing and fine-tuning, and lastly tried to provide additional data using RAG. We evaluated our final results with the automated BERTScore metric and manual human assessment to compare the different approaches and find the best model for our use case. We were able to greatly improve the base model generations and achieve reports that were more informative than the ground truth in a quarter of the cases.

## Keywords

Traffic news generation, Prompt engineering, Supervised fine-tuning, Natural language generation

Advisor: Slavko Žitnik

## Introduction

Generating traffic news using the modern Natural Language Processing (NLP) techniques involves transforming raw traffic data into structured reports that are in a form suitable for broadcasting on radio. Our main goal is to reduce manual workload for editors, and to improve the quality and consistency of prepared reports.

There are a few challenges that arise, mainly in following the formatting rules and correctly reflecting traffic conditions (e.g. road names, directions) - this is what the model has to learn to replicate. To address these conditions, we will use pre-trained language models, structured prompting, fine-tuning, and retrieval-augmented generation. We combine all these techniques together in order to be able to automatically generate high-quality Slovenian traffic reports.

## Related Work

We looked at the popular approaches for text style generation adaptations for language models. The two most prominent techniques are *prompting* and *fine-tuning*. Both try to achieve better desired results either by guiding the model to generate data according to determined rules or to change the models structure in accordance with the desired outcome. Another popular technique is retrieval-augmented generation (RAG) that allows the model to refer to a specified set of documents

to further guide and improve the response. Although this approach is more often used on larger models it could also be used to improve performance of smaller models [1].

The available research suggests that *prompting* is overall less efficient than *fine-tuning* [2] but should not be discarded. Using the right prompt can substantially increase the model's performance [3]. Additionally, smaller and medium-sized models benefit more from prompting than from simply increasing the model size [2]. The prompt should be crafted according to the desired result [4], so it is important to not rush the process of finding the best performing one.

If *prompting* is enhanced by using a *fine-tuned* model, we can achieve the best possible results. For our use case we used *fine-tuning* to make the model adapt the generated content style to the form that is used on RTV. The problem with *fine-tuning* is lack of tuning data and computing power limitations [5]. But apart from those problems the approach can heavily change and customize the generation style with noticeably better results [6].

Attempts at creating an automated traffic incident reporting were already made [7] so looking at the findings and shortcomings of existing systems will allow us to even further improve the results. Since the traffic data and the desired report structure vary greatly even between news channels, let alone countries, it is most likely impossible to create an out-of-the box solution for news reporting.

## Traffic News Dataset

In this section we present the provided dataset and the pre-processing steps we applied to the dataset to make it more suitable for language models. The dataset consists of two parts. The first part is a table that was retrieved from the Slovenian traffic website **promet.si** [8] which provides real-time traffic updates and road conditions. The second part is a corpus of traffic reports that were manually written to be read on the radio broadcast. Both parts span data from 2022 to 2024. Additionally, a set of instructions which outline the expected format of generated traffic news reports are included in the dataset to help with consistency.

### Tabular Traffic Updates Data

Traffic updates from **promet.si** are stored in a structured Excel table. It contains separate columns for different categories, such as accidents, traffic jams, roadworks, and obstacles. There are also two columns where this data is aggregated, one containing the urgent traffic updates and one containing general traffic updates.

Not much preprocessing was required on this table, as the data is already structured. Nonetheless, two important preprocessing steps had to be applied. Firstly, the entries were *converted from HTML* to a raw textual representation which is more suitable for LLMs. Secondly, we found that the timestamps in the table were given in the *UTC timezone*. This observation was made when we closely inspected two wrong-way driver reports. We localized these timestamps to make the table compatible with the unstructured reports described in the next section.

### Corpus with Traffic Reports

The corpus contains 28,037 manually written traffic reports which are stored in rich text format (*rtf*) files. Generally, the first line in the report is a header containing the report type (general or urgent) and a local timestamp. The main report content is contained in the remaining lines. Only the reports for which the header was successfully parsed were kept, as obtaining the timestamp is crucial for matching the reports with the tabular data.

After parsing the reports, empty reports were removed from the dataset. Furthermore, outliers were identified by embedding each report using a SloBERTa model, fine-tuned for topic classification [9]. The average cosine similarity between the embedding of each report and the other report embeddings was computed. The reports with the lowest average similarities were manually inspected. As a result, one additional report was excluded from the dataset, since it was not a traffic report.

### Matching the Reports with Tabular Data

As a final data preparation step, we matched the parsed reports with the entries in the traffic information table. To each report, up to 10 of the most recent entries that occurred up to 30

Preparation Stage	Number of Reports
Initial	28,037
Parsing Reports	27,567
Identifying Outliers	27,561
Matching with Tabular Data	25,415

**Table 1.** Size of the corpus after individual data preparation stages.

*minutes* before the report timestamp were assigned. Since some of these entries were identical, duplicated were removed.

For some reports, no entries were found in the table in the defined time window of 30 minutes. These reports were excluded from further analysis. An overview of corpus size after each data preprocessing step is shown in Table 1.

### Report Writing Guidelines

The dataset also contains an auxiliary file with guidelines for writing reports. It specifies that each traffic incident we report on has to follow a structured formulation, starting with a road name and direction, followed by a cause, and lastly by the consequence of the incident. It defines the hierarchy of events that need to be reported.

There are a few other rules and naming conventions that are more specific and are dependent on the type of event we are reporting on (high-priority events must be repeated every 15-20 minutes, congestion is reported only if it exceeds 1 km in length, indicate road sections with expected closures or delays). Note that some of the naming conventions for reports differ from naming conventions on *promet.si*, where the tabular data was retrieved from. This will need to be taken into account in our analysis.

## Methodology

Our goal was to investigate how structured prompting, parameter efficient fine-tuning, and retrieval can each improve the automatic generation of traffic reports. Starting with the pre-trained `cjvt/GaMS-9B-Instruct` model, we designed four distinct experimental configurations:

1. **Base instructed (prompting only)**
2. **Fine-tuned (LoRA)**
3. **Fine-tuned + instructed**
4. **Fine-tuned + instructed + mini-RAG**

All variants of our setup use the same 25,415 prompt response pairs (after deduplication and random shuffling). They were derived from our processed traffic dataset, which was split 80%/20% into training and validation sets.

### 1. Base instructed (prompting only)

We use the pre-trained `GaMS-9B-Instruct` model without any weight updates. At inference, each raw traffic prompt is enhanced with our human-readable instruction template, which enforces the proper format and naming conventions: this helps us to establish a prompting-only baseline, which we aim to improve with more sophisticated approaches.

## 2. Fine-tuned (LoRA)

To specialize the model to our traffic-report domain, we perform parameter-efficient fine-tuning with LoRA adapters on a single H100 GPU using the Arnes HPC cluster. The entire job (using 1 GPU, 8 CPU cores) took approximately 7 hours to complete.

**Training data** We fine-tune on our prompt–response examples (80% train / 20% validation) drawn from prepared dataset, where each is formatted as:

```
### Human: <raw prompt>
### Assistant: <expected report>
```

We shuffle the dataset for randomness and tokenize to a fixed length of 512 tokens, masking all “Human” tokens, so that the loss is computed only on the “Assistant” portion.

**LoRA adapter configuration** We inject low-rank modules into each attention layer’s projection matrices:

- **Rank**  $r = 8$
- **Alpha**  $\alpha = 32$
- **Dropout**  $p = 0.05$
- **Target modules:** q-proj, k-proj, v-proj, o-proj

### Training hyperparameters

- **Per-GPU batch size:** 1; **gradient accumulation:** 8
- **Learning rate:**  $2 \times 10^{-4}$  with cosine decay, warmup ratio = 0.03
- **Epochs:** 3; **Precision:** bf16

At the end, we save the LoRA adapter and tokenizer for downstream inference.

## 3. Fine-tuned + instructed

In this variant, we reuse the fine-tuned LoRA model but again add our structured instruction template at inference. This combines the benefits of domain adaptation (which happens via LoRA fine-tuning) with explicit formatting rules (via prompting).

## 4. Fine-tuned + instructed + mini-RAG

To further boost factual correctness (especially road names, sections, and formatting conventions), we enhanced the fine-tuned + instructed model with a simple Retrieval-Augmented Generation (mini-RAG) component. Our RAG pipeline includes two indexed corpora and a simple cosine-similarity retriever:

- **Road network index:** We extracted 2,500 entries from OpenStreetMap<sup>1</sup>, retaining highways, motorways, primary and secondary roads for Slovenia. Each snippet is a short textual description of a road segment, e.g.: “Dunajska cesta pelje od Križišče na (46.0681, 14.50999) do Križišče na (46.068, 14.50996). Povezuje

se s cestami: Samova ulica, Topniška ulica.”.

Metadata such as endpoints, connected roads, and type are embedded in the text. We encoded text with LaBSE and saved obtained embeddings alongside the raw text data for the purposes of retrieval.

- **Instruction rules index:** From our given writing guidelines we prepared 20 additional examples of Slovenian road-naming and formatting rules (e.g. “PRIMORSKA AVTOCESTA proti Kopru: ...”). These entries were also embedded via LaBSE with corresponding texts.

**Retrieval at inference** For each new prompt:

1. We compute the LaBSE embedding of the raw prompt.
2. We calculate cosine similarity against both the road-network and instruction rule embeddings, and we select only the top-1 match from each.
3. We add a “Retrieved documents” block listing:

1. <top road snippet>
2. <top instruction snippet>

4. We then append the original prompt and start generation with the fine-tuned + instructed model.

This mini-RAG step is supposed to add contextually relevant road details and formatting reminders, which could potentially improve the model’s ability to stick to strict naming conventions and report structure.

## Model evaluation

We applied all four approaches to generate traffic reports for a held-out set of 500 prompts, and then assessed their quality through both human evaluation and an automatic similarity metric.

### Manual evaluation

To capture aspects of readability, factual correctness, and following style guidelines that are not easily quantified, we developed a Streamlit web app allowing each of the three authors to independently rate a random sample of 30 prompts per scenario. For each example, we assigned:

- A *quality score* from 1 (poor or incorrect) to 5 (fully correct and well-formed).
- A binary indicator of whether the model’s output was *better than the original human-written report*.

Prior to the final evaluation, we calibrated our judgments on an additional 10 “warm-up” examples to align our scoring criteria. The averaged results are shown in Table 3.

Overall, the fine-tuned models substantially outperform the prompt-only baseline in human judgments. Adding structured instructions to the fine-tuned model gives us the biggest improvement, as it yields the highest average rating (3.19).

<sup>1</sup><https://download.geofabrik.de/europe/slovenia.html>

**Table 2.** BERTscore evaluation and relative report length difference of different models.

Model	Precision	Recall	F1-score	Relative length difference
Base instructed	$0.608 \pm 0.004$	$0.683 \pm 0.004$	$0.643 \pm 0.004$	$1.904 \pm 0.042$
Fine-tuned	$0.774 \pm 0.003$	$0.753 \pm 0.004$	$0.762 \pm 0.003$	$0.818 \pm 0.022$
Fine-tuned and instructed	$0.817 \pm 0.003$	$0.752 \pm 0.004$	$0.781 \pm 0.003$	$0.732 \pm 0.017$
Fine-tuned and instructed + RAG	$0.815 \pm 0.003$	$0.752 \pm 0.004$	$0.779 \pm 0.003$	$0.752 \pm 0.018$

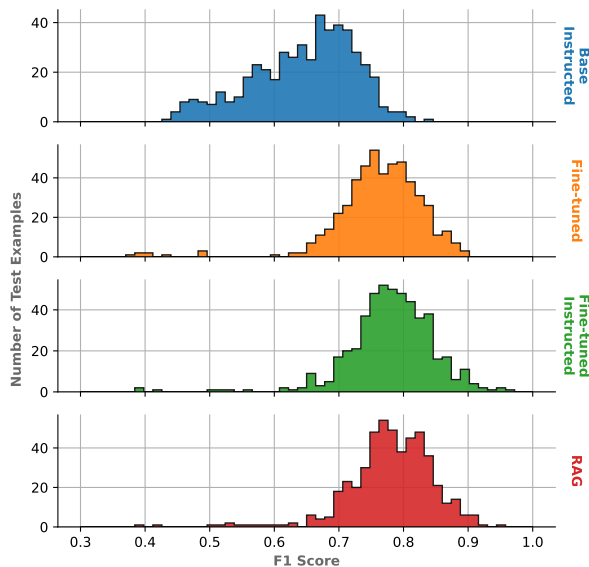
**Table 3.** Manual evaluation of different models.

Model	Rating	Beats GT [%]
Base instructed	$1.97 \pm 0.10$	$6.7 \pm 2.6$
Fine-tuned	$3.01 \pm 0.13$	$22.2 \pm 4.4$
Fine-tuned + instr.	$3.20 \pm 0.11$	$28.9 \pm 4.8$
Fine-tuned + instr. + RAG	$3.07 \pm 0.12$	$25.6 \pm 4.6$

### Automatic evaluation

To get an objective measure, we computed BERTscore (using the SloBERTa model [10]) precision, recall, and F1 between each generated report and its reference. Due to compute constraints, this evaluation covered 500 examples per model. We also measured the *relative length difference* to ensure consistency. Results are summarized in Table 2 and the full F1 distributions appear in Figure 1.

The downside of this method is that it is not able to directly check if the structure of a report is suitable, or whether the reported events in fact occurred, and it can not explicitly check if reporting conventions were taken into account. Furthermore, as observed in the data exploration phase, some ground-truth reports do not contain the latest traffic information, which might cause our models to be underestimated. That is why we combined both manual and automatic evaluation approaches to holistically benchmark the models.

**Figure 1.** Distributions of F1 scores of all tested models.

### Length consistency

We looked at how the generated report lengths coincide with the ground truth length. Since the distribution of report length for the ground truth is centred around 80 words, we wanted the model to generate reports of similar length. To do this, we computed the relative difference between the number of words in the generated and the ground truth reports. The results are shown in Table 2 along with the BERTscore results. The base instructed model tends to generate reports that are too long, on average being almost two times longer than the ground truth. The fine-tuned models perform better in this aspect.

## Discussion and Results

### Data quality

Our work was frequently challenged by noisy and inconsistent input data. The structured traffic updates from *promet.si* often violated naming and formatting rules, and timestamps appeared interchangeably in UTC and local time. We also encountered non-traffic entries such as weather reports or unrelated news items. On the output side, the human-written “ground truth” reports themselves lacked uniformity: some skipped critical incidents flagged in the input, while others included information not present in the structured data. This mismatch complicated both training and evaluation, and required careful preprocessing to mitigate its impact.

### Final performance

Across all metrics, our fine-tuned variants outperformed the prompt-only baseline. Table 3 shows that simple LoRA fine-tuning boosted human-rated quality from 1.97 to 3.01, and the addition of structured instructions further raised the average to 3.20. Introducing mini-RAG produced a slight dip to 3.07, which means that our small retrieval corpus did not meaningfully enhance output quality.

Similarly, the automatic SloBERTa-based evaluation confirms these trends: Precision, Recall, and F1 for the fine-tuned + instructed variant had a high increase from the base model. Figure 1 further illustrates that fine-tuning and instructions also produce a tighter F1-score distribution, which means more consistent outputs.

Our current mini-RAG setup yielded limited benefit, which could be due to the small size of our retrieval corpus or the model’s capacity constraints. Future work could explore richer retrieval sources to improve the use of external road-network and rule-based knowledge.

## References

- [1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.
- [2] Martina Toshevskaja and Sonja Gievska. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721, 2025.
- [3] Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. Does prompt formatting have any impact on llm performance?, 2024.
- [4] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. Why johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI ’23, New York, NY, USA, 2023. Association for Computing Machinery.
- [5] Yuchen Xia, Jiho Kim, Yuhang Chen, Haojie Ye, Souvik Kundu, Cong Callie Hao, and Nishil Talati. Understanding the performance and estimating the cost of llm fine-tuning. In *2024 IEEE International Symposium on Workload Characterization (IISWC)*, pages 210–223, 2024.
- [6] Xinyue Liu, Harshita Diddee, and Daphne Ippolito. Customizing large language model generation style using parameter-efficient finetuning, 2024.
- [7] Anna Jansdatter Bjørge and Mads Lundegaard. Automating the process of traffic incident reporting. Master’s thesis, NTNU, 2024.
- [8] DARS d.d. Prometno informacijski center za državne ceste. <https://www.promet.si>. Accessed: 20. 03. 2025.
- [9] CJVT. Sloberta for text topic classification. <https://huggingface.co/cjvt/sloberta-trendi-topics>. Accessed: 05. 02. 2025.
- [10] Matej Ulčar and Marko Robnik-Šikonja. Slovenian roberta contextual embeddings model: Sloberta 2.0. *clarin.si*, 2021.

## Appendix

### Example of a well generated report

This example shows generated report that aligns well with the ground-truth. Furthermore, it contains additional, arguably important information about fog that is not included in the ground-truth report.

#### Model Input

##### *Nesreče*

- Na južni ljubljanski obvoznici sta zaprta vozni in prehitevalni pas na razcepu Kozarje iz smeri Viča proti Brezovici. Nastaja zastoje.

##### *Ovire*

- Cesta Vrhnika - Horjul - Šentjošt nad Horjulom je pri Vrzdencu zaprta zaradi zemeljskega plazu in podrtega drevesa. Obvoz je možen preko Polhovega Gradca.

##### *Delo na cesti*

- Ljubljanska severna obvoznica bo zaprta med Zadobrovo in Novimi Jaršami v obeh smereh v noči na nedeljo od 22. do 13. ure.
- Cesta Brezovica - Podpeč bo v Notranjih Goricah zaprta danes do 22. ure.

##### *Vreme*

- Zaradi megle je ponekod po Sloveniji zmanjšana vidljivost. Previdno!

##### *Zimske razmere*

- Cesta čez prelaz Vršič je zaradi zimskih razmer zaprta.

#### Ground-Truth

Podatki o prometu.

Zaradi prometne nesreče sta ljubljanski južni obvoznici na razcepu Kozarje zaprta vozni in prehitevalni pas iz smeri Viča proti Brezovici.

Cesta Vrhnika-Horjul-Šentjošt nad Horjulom je zaradi zemeljskega plazu zaprta pri Vrzdencu. Obvoz je prek Polhovega Gradca.

#### Model Output

Podatki o prometu.

Na južni ljubljanski obvoznici sta zaradi prometne nesreče na razcepu Kozarje zaprta vozni in prehitevalni pas iz smeri Viča proti Brezovici.

Cesta Vrhnika-Horjul-Šentjošt nad Horjulom je zaradi zemeljskega plazu zaprta pri Vrzdencu. Obvoz je prek Polhovega Gradca.

Ponekod po državi je zaradi megle zmanjšana vidljivost.

### Example where the generated report contains information that is no longer relevant

In the example below, the report generated by the model contains information about the lane closure on south Ljubljana bypass, which is no longer relevant. This is likely due to the fact that the input data contains entries from the 30 minutes prior to the report – some of which may no longer be relevant.

#### Model Input

##### *Nesreče*

- Na štajerski avtocesti je zaprt vozni pas med priključkom Celje zahod in Celje center proti Mariboru. Nastaja zastoje.

##### *Ovire*

- Na južni ljubljanski obvoznici je pred priključkom Ljubljana jug proti Malencam zaprt vozni pas, okvara vozila. Nastaja zastoje.

##### *Tovorna vozila*

- Omejitev za tovorna vozila, katerih največja dovoljena masa presega 7,5 t v času novoletnih praznikov je med 8. in 22. uro v sredo in četrtek, 1. in 2. januarja 2025.

#### Ground-Truth

Podatki o prometu.

Na štajerski avtocesti je zaradi nesreče zaprt vozni pas med priključkom Celje-zahod in Celje-center proti Mariboru.

#### Model Output

Podatki o prometu.

Na štajerski avtocesti proti Mariboru je zaradi nesreče zaprt vozni pas med priključkom Celje-zahod in Celje-center.

Na južni ljubljanski obvoznici proti Dolenjski je zaradi pokvarjenega vozila zaprt vozni pas pred priključkom Ljubljana-jug. Nastal je krajši zastoje.

### Example of hallucinations

#### Model Input

##### *Vremenske razmere*

- Na cesti Zagorje - Podkum pri Šklendrovcu zaradi zemeljskega plazu promet poteka izmenično enosmerno.

##### *Dela na avtocestah*

- Na primorski avtocesti bo do 24. junija zaprt izvoz Vrhnika iz smeri Kopra. Obvoz je preko priključka Logatec.
- V petek, 3. junija, od 15:30, pa bodo zaprli še uvoz proti Ljubljani. Zapore bo do nedelje do 7. ure.

##### *Zapore na glavnih in regionalnih cestah*

- Zaprta je cesta Žiče - Slovenske Konjice, v Žičah, do 19. 6. Cesta Železniki - Kropa bo pri Lajšah zaprta do 15. 6. zaradi sanacije plazu.
- Cesta Novo mesto - Metlika bo pri Dolnji Lokvici zaprta do 3. 6. 2022.

#### Ground-Truth

Podatki o prometu.

Zaradi del je na primorski avtocesti proti Ljubljani zaprt izvoz Vrhnika. Obvoz je prek priključka Logatec.

Na regionalni cesti Zagorje-Podkum je zaradi zemeljskega plazu promet urejen izmenično enosmerno pri Šklendrovcu.

#### Model Output

Podatki o prometu.

Na primorski avtocesti proti Ljubljani je zaradi del zaprt izvoz Vrhnika. Obvoz je prek priključka Logatec.

Prav tako zaradi del je na štajerski avtocesti proti Mariboru zaprt vozni pas med priključkom Celje-zahod in Celje-center.