Verlässlichkeit domänenspezifischer Sprache-nach-Text-Umwandlung am Beispiel des digitalisierten Zugleitbetriebs

Individual Research Project

Betreut von Lukas Pirl

Luca Kleinschmidt

21.07.2026

Recap: Idee

- Verlässlichkeit in Machine-Learning-Komponenten des ZLiC
- Ausgabe der Sprache-zu-Text entlang einer Grammatik
- Vergleichbarkeit für Abstimmungen zwischen verschiedenen Modellen und Grammatikstrategien

Recap: Motivation

- Verwendete Modelle sind inhärent nicht erklärbar
- Zuverlässigkeit des Sprache-zu-Text so *weit wie möglich* erhöhen, dass das System meistens in funktionalem Zustand ist.
- Schnellere Reaktionszeit des Systems durch neuere Technologien
- Funktionale Sicherheit durch Funkprotokoll des Betriebsverfahrens[1]

ZLiC Baseline

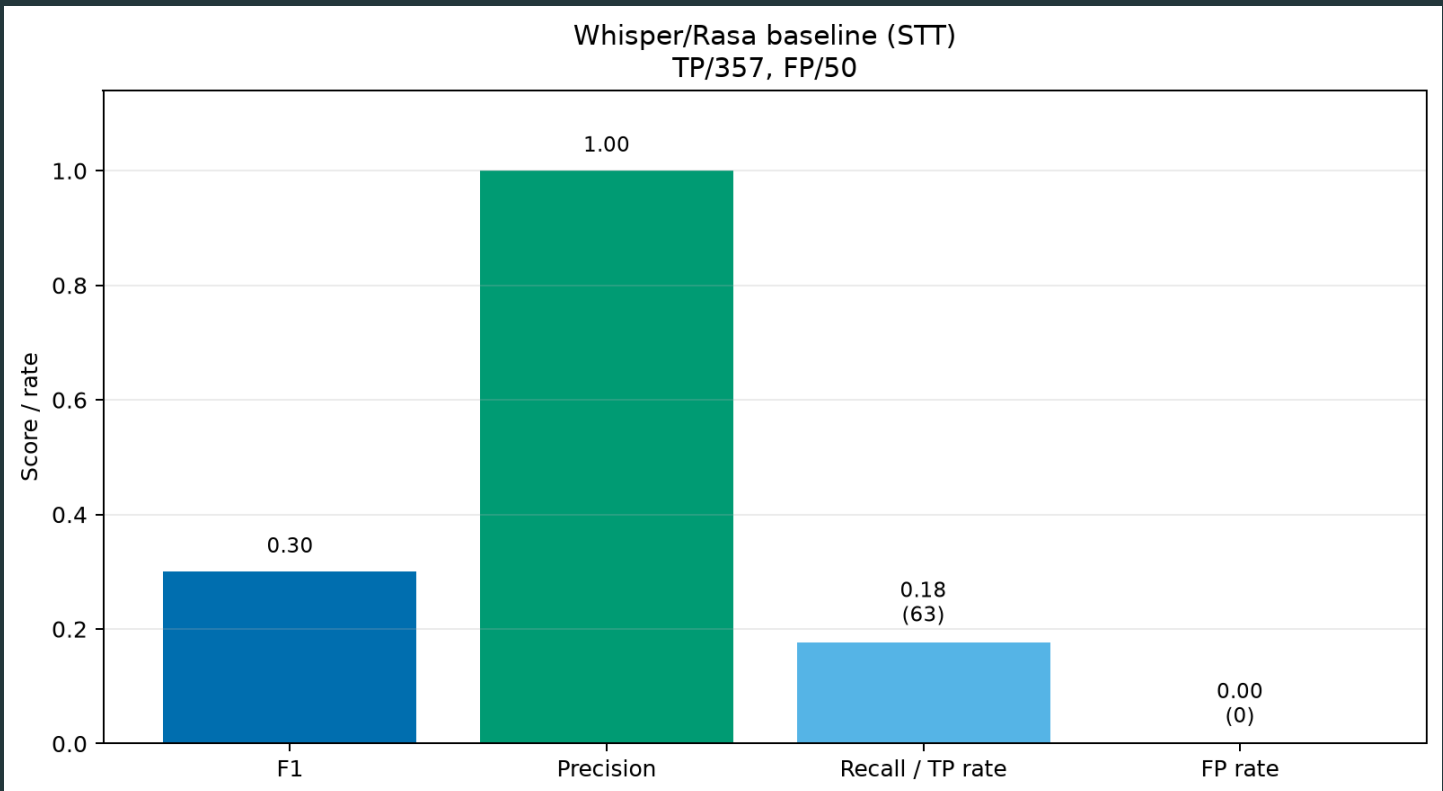


Figure 1: ZLiC Paper-Implementierung mit Whisper und Rasa [1]

- Recall erhöhen aber wie?

Recap: Technische Details ZLiC Umbau

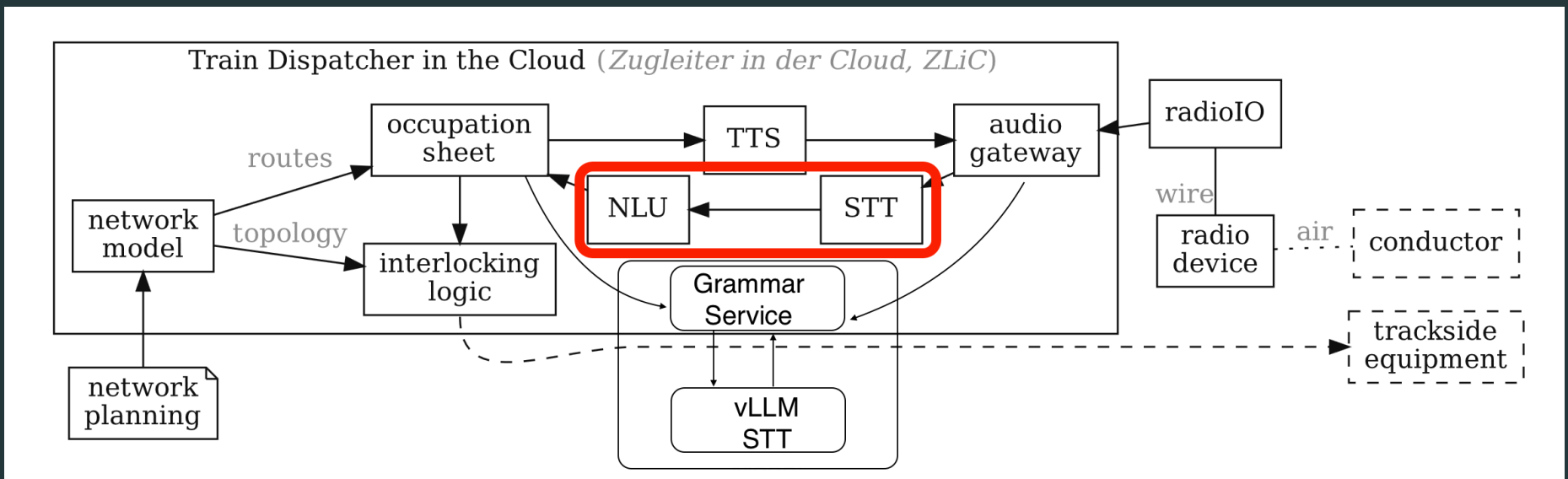


Figure 2: Architekturdiagramm des “Zugleiter in der Cloud”[1] Projekts

Modelle: “Mistral Voxtral Mini” [2] und “Google Gemma 4 E2B” [3] mit vLLM [4] und Xgrammar [5]

Recap: Grammatik

- Garantiert syntaktisch korrekte Ausgabe auch bei *jeder* Eingabe [6]
- Semantik der Eingabe unterscheidet sich ggf. von der von der Grammatik akzeptierten Semantik [7]
- Wenn sich die Eingabesemantik von der akzeptierten Semantik nicht unterscheidet: Grammatik “führt” zu syntaktisch korrekten Ausgaben
- Token-Triggered-Grammar: Grammatik erst bei Beginn der Syntax [8]

Falsch Positive Ergebnisse

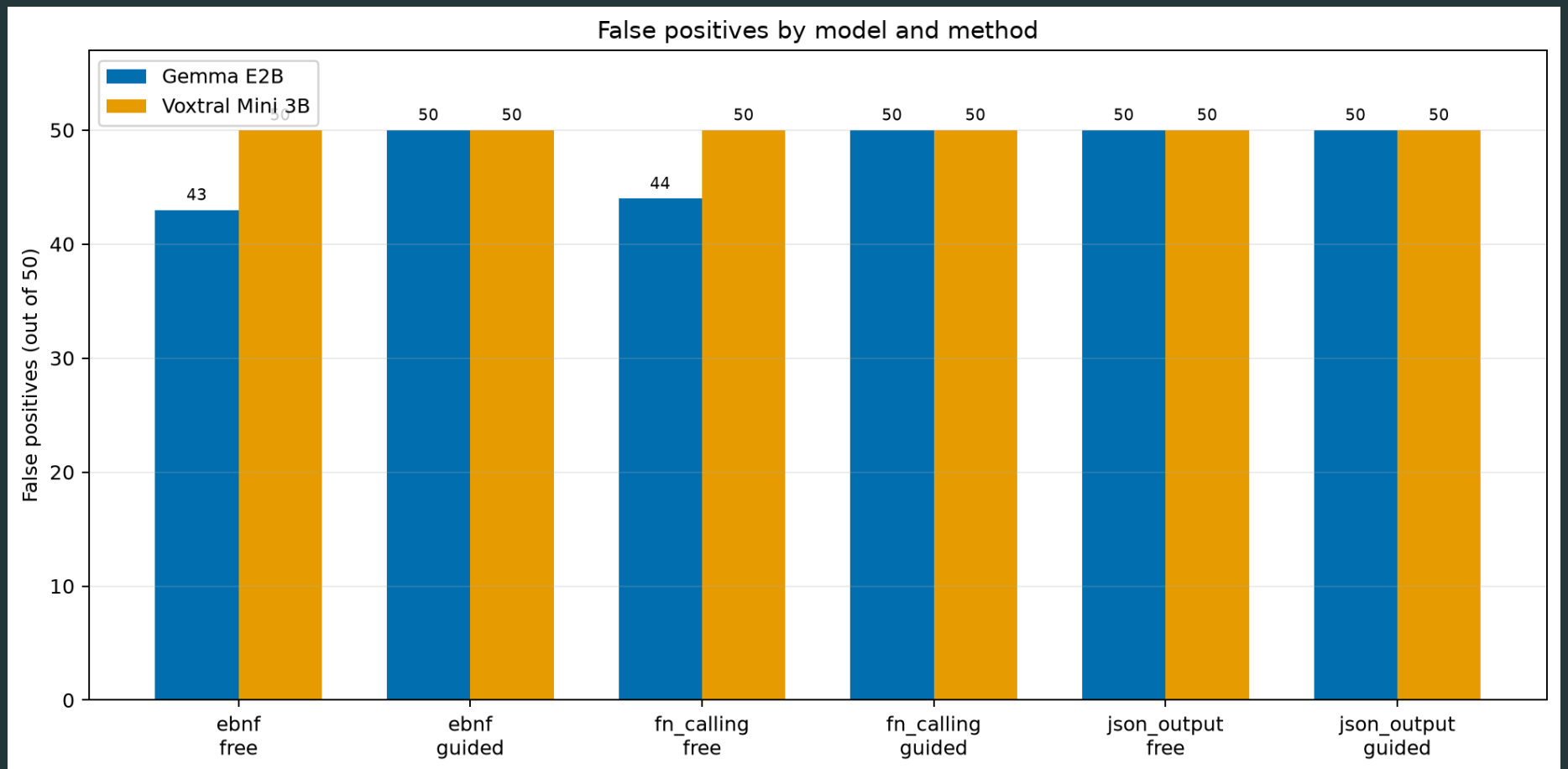


Figure 3: Abstimmung mit einfacher Mehrheit zwischen Voxtral Mini [2] und Gemma 4 E2B [3]

Probleme mit klassischen Abstimmungen bei Maschinellem Lernen

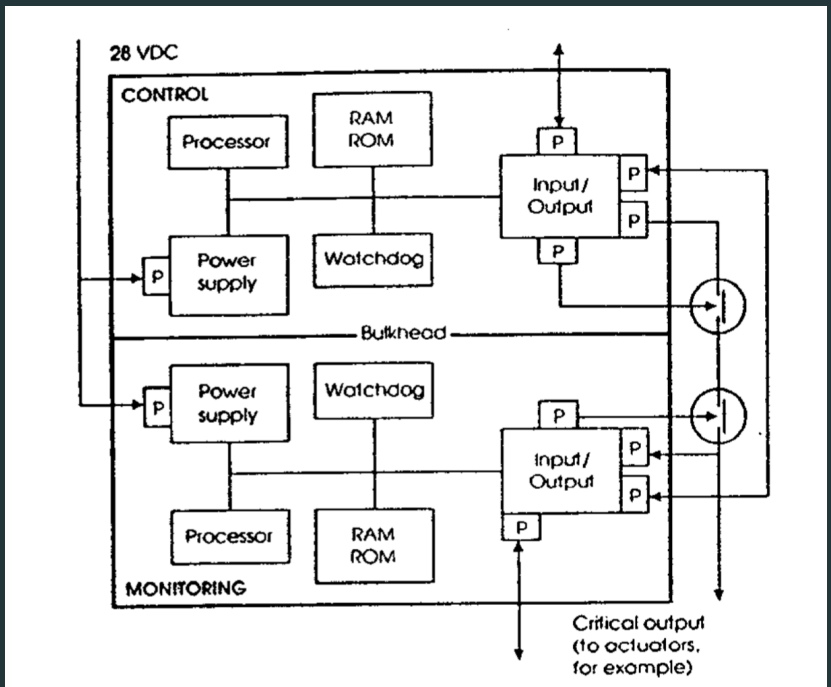


Figure 4: Flugkontrollcomputerarchitektur des Airbus A320/A330/A340 [9] mittels Wachhund

- Wachhundprinzip
- Zwei unabhängige Implementierungen mit z.B. Intel i386 und Motorola m68k
- Annahme in 99.9% der Fälle gleiche Ergebnisse

Übertragung auf Maschinelles Lernen

- Viele Ergebnisse mit hoher Wahrscheinlichkeit unterschiedlich
- Behandlung von Ungenauigkeiten in Modellausgaben
- Behandlung von gleichartigen Fehlerursachen
- Verschiedene ResNet-Modellgrößen mit TMR gegen Hardwarefehler [10]
- Robustheit des Modells durch Transformation (data augmentation) der Eingabe [11]

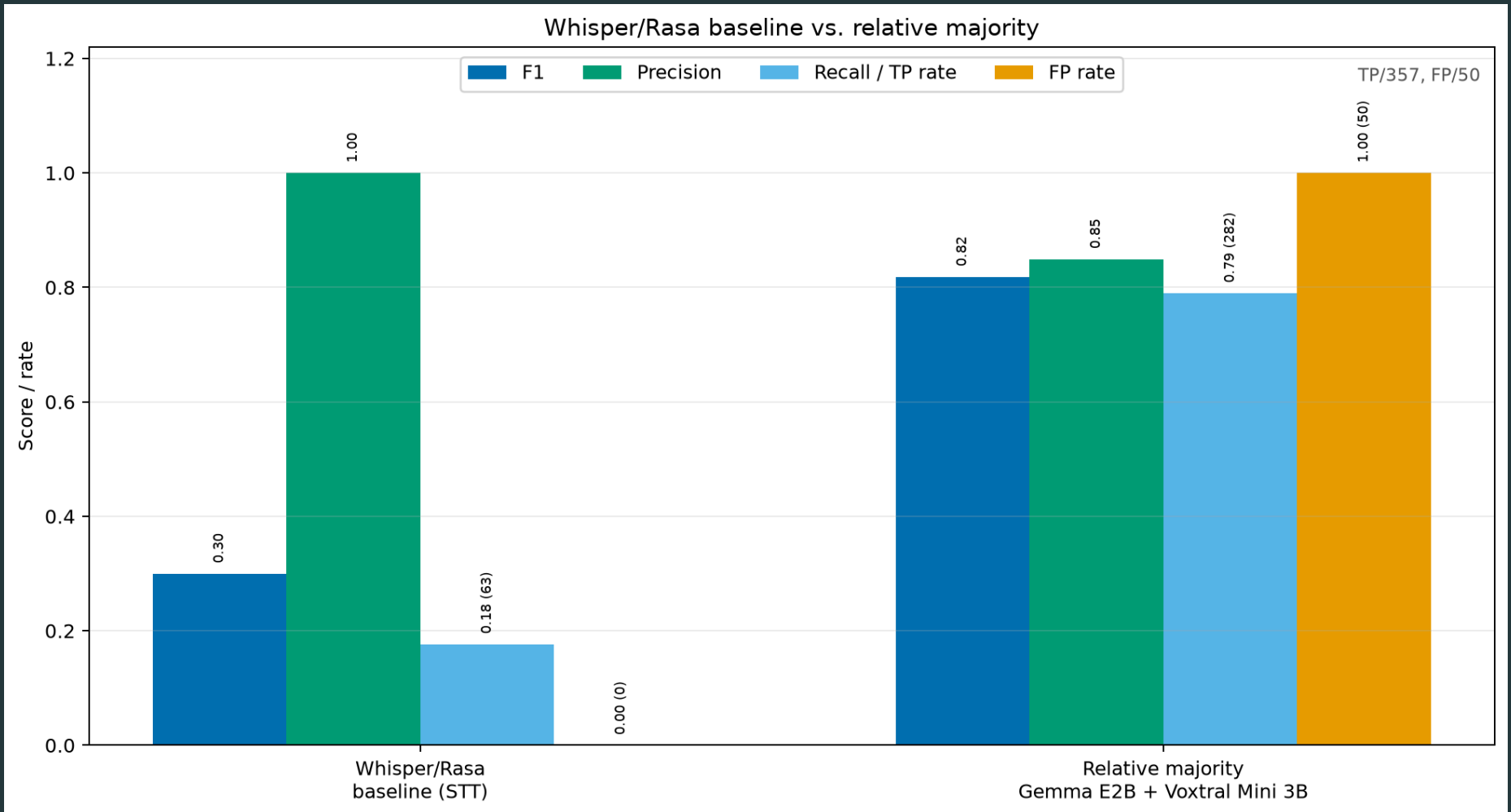


Figure 5: Abstimmung mit einfacher Mehrheit zwischen Voxtral Mini [2] und Gemma 4 E2B [3]

Dimensionen

- Autoregressive Modelle schwer vergleichbar auf Token-Ebene
- Vergleichbare Ergebnisse durch Diversität:
  - Modellarchitektur/Implementierung/Training/Größe
  - Strategien der Anfrage/Ausgabe

Beispiel

- Drei Modelle M_χ und drei Grammatiken G_χ

$$(M_a, G_\delta), (M_a, G_\varepsilon), \left(M_a, G_\varphi\right), (M_b, G_\delta), (M_b, G_\varepsilon), \left(M_b, G_\varphi\right), (M_c, G_\delta), (M_c, G_\varepsilon), \left(M_c, G_\varphi\right).$$

Dreifache modulare Redundanz entlang zwei Dimensionen

- dimensionen! Einheiten auf der ersten Ebene
- Steigende Komplexität mit Dimensionen, um Einigkeit in jeder Dimension zu sichern

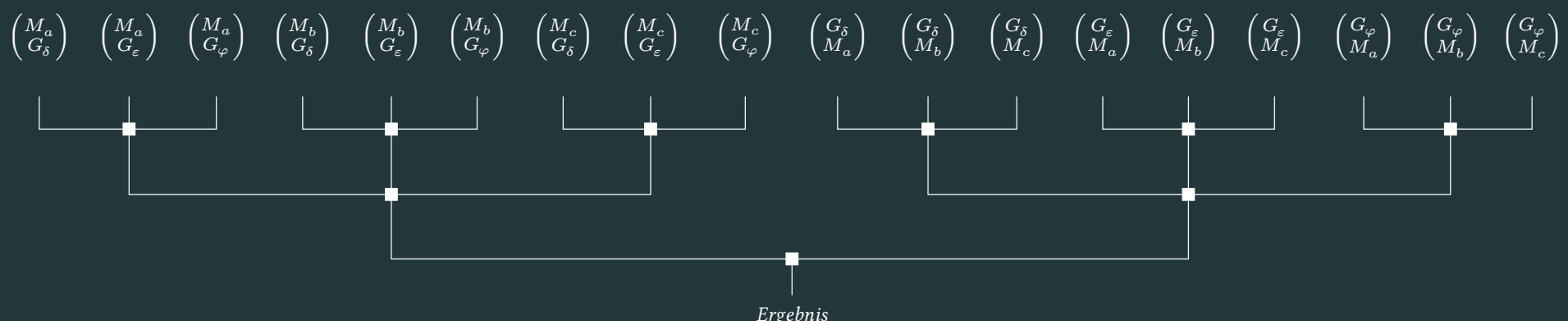


Figure 6: Hierarchische Voting-Struktur

Einfachere Vergleichbarkeit schaffen

- Diversität gleicher Antworten quantifizieren
- Bestrafung von gleichartigen Fehlerursachen
- Intuitiv: Wenig diverse Ergebnisse sind besser als viele ähnlicher Implementierungen
- Festlegung von Mindestanforderungen
- Anforderungen an Abstimmungsstrategien
  - Falsche Positive vermeiden
  - Möglichst viele richtige Positive behalten

Beispiel: Multi-Hot-Vektoren

n-dimensionale Vektoren Positionsvektor $\begin{pmatrix} M_a \\ M_b \\ M_c \\ G_\delta \\ G_\varepsilon \\ G_\varphi \end{pmatrix} \rightarrow \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$

Metrik

- Volumen auf der Kovarianz-Matrix der Multi-Hot-Vektoren
- Intuition: Volumen und wie viele Richtungen
- Isolierung gemeinsamer Komponenten der Multi-Hot-Vektoren

Beispiele Diversität

Viel gemeinsamer Fehler bei 2 Kategorien mit je 3 Möglichkeiten

$$S\left(\begin{pmatrix}1&0&0&0&0&1\\1&0&0&0&1&0\\1&0&0&1&0&0\end{pmatrix}\right) < S\left(\begin{pmatrix}0&0&1&0&0&1\\0&1&0&0&1&0\end{pmatrix}\right)$$

Wenig gemeinsamer Fehler bei 3 Kategorien mit je 2 Möglichkeiten

$$S\left(\begin{pmatrix}1&0&1&0&0&1\\0&1&1&0&1&0\end{pmatrix}\right) < S\left(\begin{pmatrix}0&1&0&1&1&0\\1&0&1&0&0&1\\0&1&0&1&0&1\end{pmatrix}\right)$$

Algorithmus

1. Gruppieren nach gleichen Ergebnissen
2. Gruppen aussortieren, die Voraussetzungen nicht erfüllen
   - Mindestens 2 gleiche Ergebnisse in einer Gruppe
   - Gesetzte Mindestanforderungen z.B. mindestens 2 Modelle
3. Ergebnis der Gruppe $\text{argmax}_{g \in G} S(g)$ zurückgeben

Abstimmung mit Diversitätsmetrik und Score

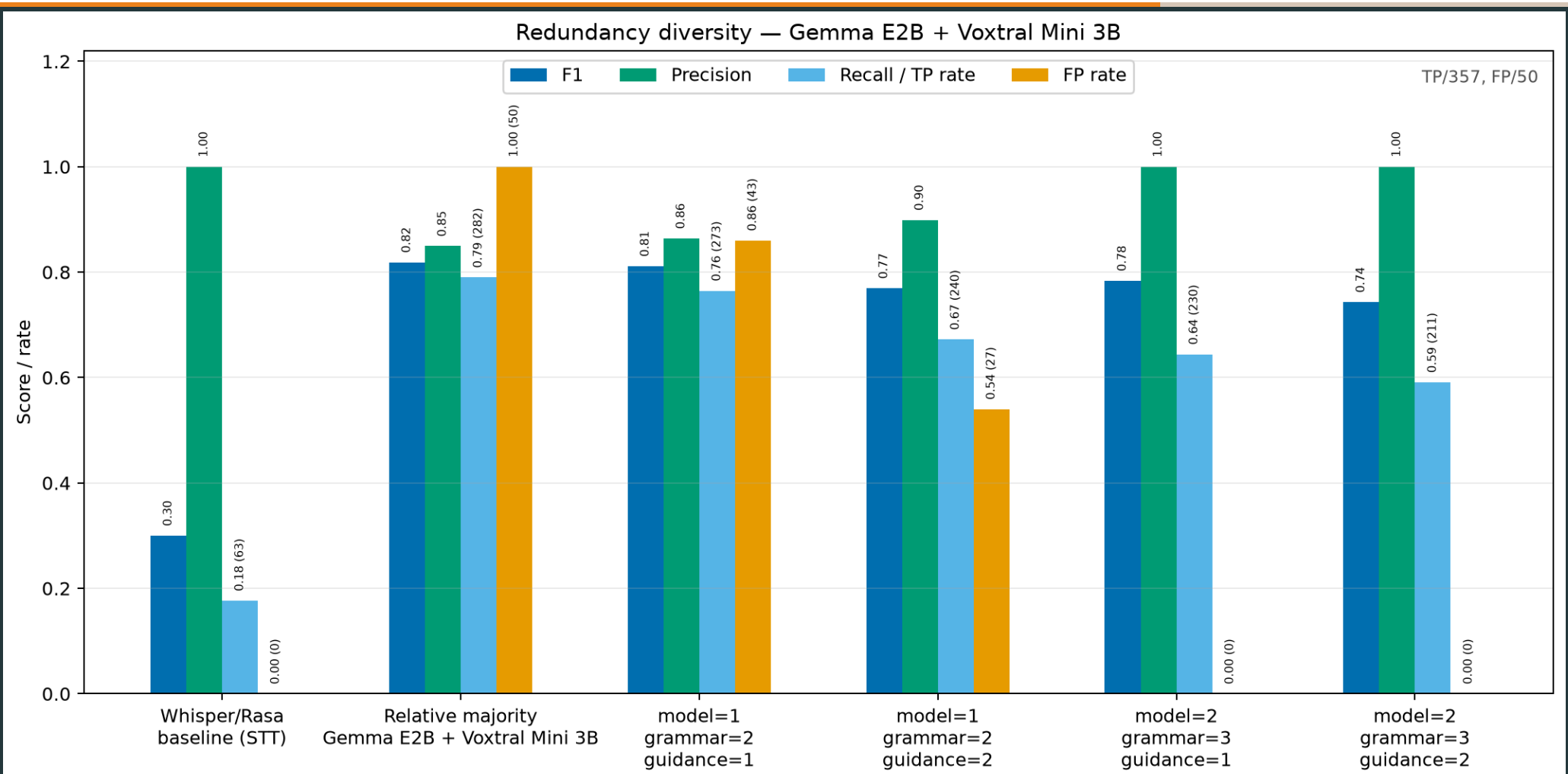


Figure 7: Abstimmung mit einfacher harten Voraussetzungen, Diversitätsmetrik und Score

Verhalten bei leistungsfähigeren Modellen

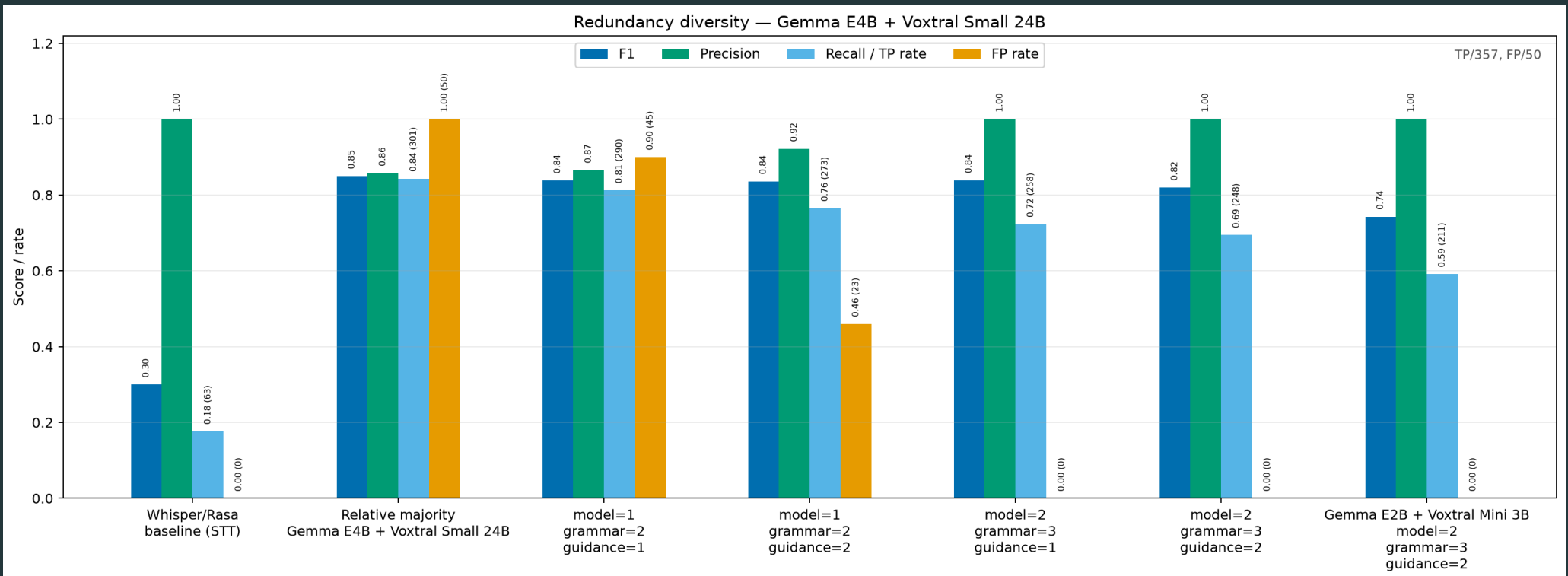


Figure 8: Größere Modelle mehr besser :)

Einfluss des Scores auf die Abstimmung

Kombination	Konfiguration	Score TP	Score FP	Mehrheit mit Kriterium TP	Mehrheit mit Kriterium FP	ΔTP	ΔFP
E2B + Mini 3B	model=1, grammar=1, guidance=2	251	50	246	50	+5	+0
E2B + Mini 3B	model=1, grammar=2, guidance=1	273	43	276	47	−3	−4

Table 1: Score beeinflusst die Abstimmung

- Harte Voraussetzungen reduzieren die Anzahl der Kombinationen
- Score selten entscheidend
- Bei steigender Gruppenanzahl und Diversität in einer Gruppen bei weniger harten Voraussetzungen ist Score entscheidend
- Robustheit bei uneindeutigen oder korrelierten Fehlern
- Vermeidet Scheinkonsense

Zusammenfassung

- ZLiC auf Grammatiken umgerüstet
- Aufbau eines Testdatensatzes
- Vermessung von Modellen und Strategien
- 3 Grammatikstrategien und 2 Führungsstrategien
- Metrik zur Bewertung gleicher Antworten anhand Diversitätskriterien
- Testen der Metrik

Metriken

- Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1:

$$F_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Literatur

[1] L. Pirl, H. Herholz, D. Friedenberger, A. Boockmeyer, A. Polze, and B. Milius, “Train Dispatcher in the Cloud – Digitalising Track Warrant Control for Safe Train Operations in Structurally Transforming Areas,” in *Transport Transitions: Advancing Sustainable and Inclusive Mobility*, C. McNally, P. Carroll, B. Martinez-Pastor, B. Ghosh, M. Efthymiou, and N. Valantasis-Kanellos, Eds., Cham: Springer Nature Switzerland, 2024, pp. 13–19.

[2] A. H. Liu *et al.*, “Voxtral.” [Online]. Available: https://arxiv.org/abs/2507.13264

[3] “Gemma 4 model card.” [Online]. Available: https://ai.google.dev/gemma/docs/core/model_card_4

[4] W. Kwon *et al.*, “Efficient Memory Management for Large Language Model Serving with PagedAttention,” in *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[5] Y. Dong *et al.*, “Xgrammar: Flexible and efficient structured generation engine for large language models,” *Proceedings of Machine Learning and Systems 7*, 2024.

[6] S. Geng, M. Josifoski, M. Peyrard, and R. West, “Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, H. Bouamor, J. Pino, and K. Bali, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 10932–10952. doi: 10.18653/v1/2023.emnlp-main.674.

[7] D. Banerjee, T. Suresh, S. Ugare, S. Misailovic, and G. Singh, “CRANE: Reasoning with constrained LLM generation,” in *Proceedings of the 42nd International Conference on Machine Learning*, in Proceedings of Machine Learning Research, vol. 267. Vancouver, Canada: PMLR, 2025.

[8] L. Li, Y. Dong, G. Wang, Z. Xu, A. Jiang, and T. Chen, “XGrammar-2: Efficient Dynamic Structured Generation Engine for Agentic LLMs.” [Online]. Available: https://arxiv.org/abs/2601.04426

[9] D. Briere and P. Traverse, “AIRBUS A320/A330/A340 electrical flight controls - A family of fault-tolerant systems,” in *FTCS-23 The Twenty-Third International Symposium on Fault-Tolerant Computing*, 1993, pp. 616–623. doi: 10.1109/FTCS.1993.627364.

[10] Z. Gao *et al.*, “Soft Error Tolerant Convolutional Neural Networks on FPGAs With Ensemble Learning,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 30, no. 3, pp. 291–302, 2022, doi: 10.1109/TVLSI.2021.3138491.

Literatur

[11] M. Caro, A. Brando, and J. Abella, “Semantic Diverse DMR and TMR for High-Integrity AI-Based Function Efficiency,” *ACM Trans. Cyber-Phys. Syst.*, vol. 9, no. 2, Mar. 2025, doi: 10.1145/3716140.

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision.” [Online]. Available: https://arxiv.org/abs/2212.04356

[13] F. Raspanti, T. Ozcelebi, and M. Holenderski, “Grammar-Constrained Decoding Makes Large Language Models Better Logical Parsers,” in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, G. Rehm and Y. Li, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 485–499. doi: 10.18653/v1/2025.acl-industry.34.

[14] I. T. Jolliffe and J. Cadima, “Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.