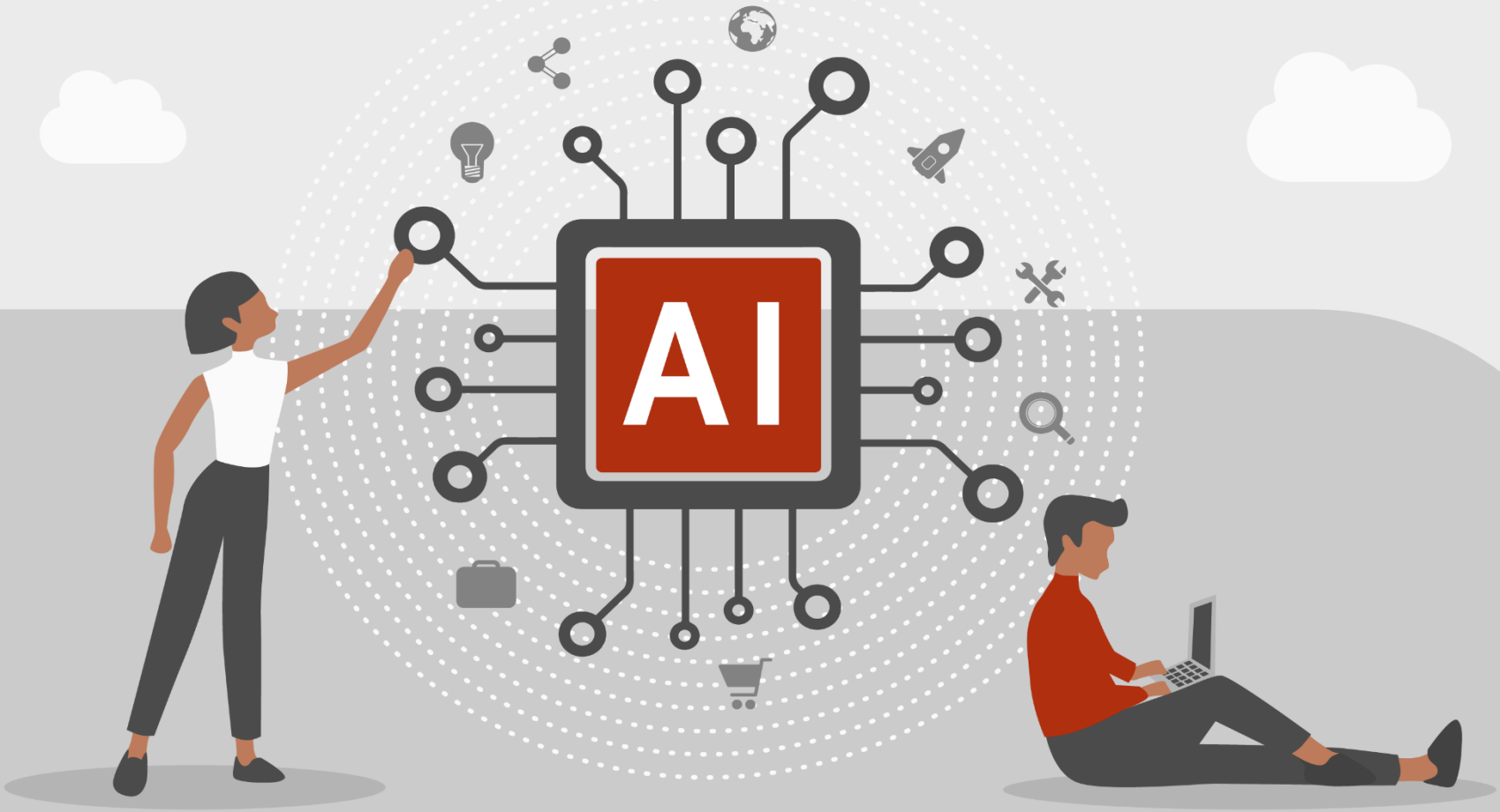
Wie komme ich zu meinem eigenen Sprachmodell?



parson

Who we are

- Gegründet von Ulrike Parson in 2006

- **Hamburg**
- Berlin, Potsdam
- Freiburg
- Hildesheim, Bremen

- 15 Technical Communicators und Consultants
- 4 Administratoren

Expert:innen für smarten Content und Informationsarchitektur

- Einführung von Informationsmanagementsystemen
- Automatisierung von Content-Prozessen
- Content-Strategie und Digitalisierung
- Technische Redaktion und Informationsarchitektur

Helle Hannken-Illjes
Technical Communicator

Ulrike Parson
CEO

Worum es heute geht

- Forschungsprojekt
- Finetuning eines Sprachmodells mit Kundendaten Technischer Dokumentation
- Sprachmodell soll beim Schreiben eines Textes das nächste Wort vorhersagen
- Ziel: Hat Finetuning mit einer begrenzten Datenmenge einen messbaren Effekt auf die Vorhersagegenauigkeit des Modells?

Quelle: Fotolia

parson

Agenda

Was sind Sprachmodelle (Large Language Models)

Anwendung von Sprachmodellen in der Technischen Kommunikation

Finetuning eines Sprachmodells: Vorgehensweise und Ergebnisse

Ausblick auf den produktiven Einsatz in der Technischen Redaktion

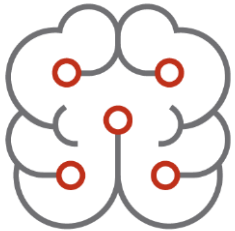
parson

Was sind Sprachmodelle?

parson

Was sind Sprachmodelle

- Künstliches Neuronales Netzwerk, das durch maschinelles Lernen Sequenzmuster natürlicher Sprache erlernt
- Sprachmodelle lösen verschiedene Aufgaben
- Sprachmodelle sind die „Gehirne“ von Anwendungen wie ChatGPT und Alexa
- Open-source (OS) und closed-source (CS) Modelle
- Bekannte Modelle, siehe auch https://huggingface.co/:
  - gpt3.5 und gpt4 von OpenAI (CS)
  - LLaMA von Meta (OS)
  - Claude von Anthropic (OS)
  - Bloom (OS)
  - PaLM von Google (CS)



parson

Merkmale von Sprachmodellen

- Technische Eigenschaften
  - Anzahl der Parameter
  - Anzahl der Layer
- Verwendete Daten
  - Inhalt
    - Korpus aus Social Media Konversationen (Tweets etc.)
    - Common Crawl
    - Geprüfte Datensätze
  - Sprache
    - Sprache der Daten
    - Besonderheiten der Sprachen

parson

Anwendung von Sprachmodellen in der Technischen Kommunikation

parson

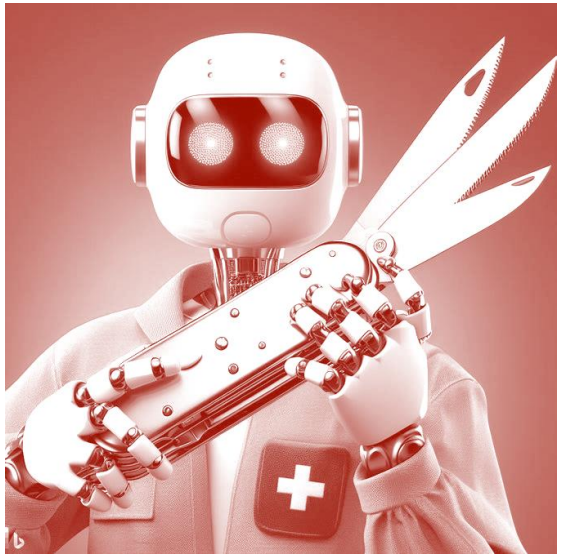
Vielfältige Anwendungsmöglichkeiten

Texte erstellen

- Stichwörter zu Text
- Fakten zu Warnhinweis
- Code erzeugen

Texte überarbeiten

- Zusammenfassen
- Weiter ausführen
- Verständlichkeit verbessern
- Sprachstil ändern
- Semantik prüfen
- SEO verbessern



Generated with Bing Image Creator

Content verwalten

- Wiederverwendbare Topics finden
- Ziele von Querverweisen finden
- Indexeinträge vorschlagen
- Metadaten zuweisen

Maschinelle Übersetzung

Content nutzen

- Chatbots im Kundenservice
- Doku im Portal befragen
- Content-Nutzung analysieren

parson

Risiken beim Einsatz

© strichfiguren.de, fotolia

- Von generativer KI erzeugte Texte sind statistische Ergebnisse und Ergebnisse von deduktiven Lernprozessen
  - Korrektheit, Zuverlässigkeit und Belastbarkeit der Aussagen nicht garantierbar.
  - Vorurteile (bias) möglich
  - Nachvollziehen von Entscheidungen aufwändig
  - Vergessen von Informationen aufwändig
  - Begrenztes domänenspezifisches Wissen
  - Keine Berücksichtigung von Rollen und Rechten
- Nutzung von Cloud-Diensten: Datenschutz und Datensicherheit sicherstellen
- Hoher Rechenaufwand für KI-Anwendungen: Ökologischer Fußabdruck und Kosten
- Know-how-Aufbau im Unternehmen
- Rechtliche Rahmenbedingungen größtenteils ungeklärt

parson

Finetuning eines Sprachmodells

Vorgehensweise und Ergebnisse

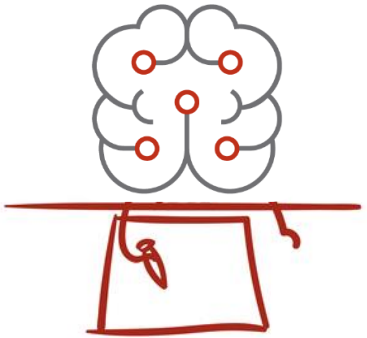
Forschungsprojekt: Sprachmodell mit eigenen Daten

- Hat das Finetuning mit Technischer Dokumentation einen messbaren Effekt auf die Qualität der Vorhersagen eines Sprachmodells?
- Projektpartner
  - parson: Anwendungsfall aus Technischer Dokumentation
  - LangTec: KI-Entwicklung und -Beratung
  - Endress+Hauser: Bereitstellung der Daten
  - SyncRO Soft: Plugin-Entwicklung für Integration in Oxygen

parson

Use Case

- Schreibunterstützung für Technische Redaktion
  - Next-word prediction bzw. next-token prediction
  - Zielgruppe: fachliche Expert:innen
  - Anwendung: beim Schreiben im Alltag
  - Ziel: Verbesserung der Konsistenz über Texte und Redakteur:innen hinweg
- Out-of-scope
  - Volle Integration in Tools
  - Tauglichkeit für produktiven Einsatz



parson

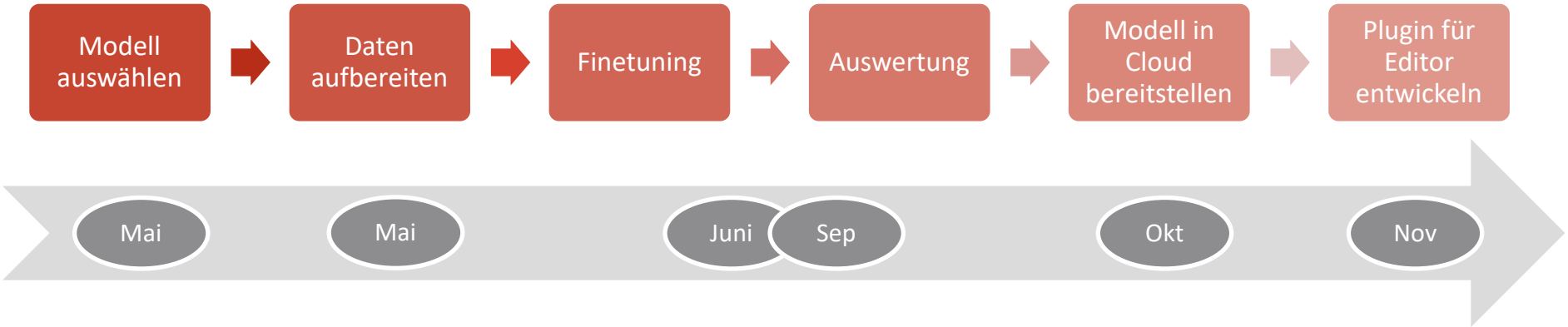
Daten für das Finetuning

- Über 2000 Dokumente der Technischen Dokumentation von Endress+Hauser
  - Jeweils auf Englisch und Deutsch
  - Bedienungsanleitung, Technische Informationen, Softwaredokumentation
    - Konzepte
    - Anleitungen
    - Listen
    - Tabellen

© strichfiguren.de, fotolia

parson

Prozessüberblick



parson

Auswahl des Sprachmodells

Modell auswählen

Kriterien für die Auswahl

- Anzahl der Parameter des Modells
- Benötigter Speicherplatz und RAM für das Finetuning
- Sprachen im Modell
- Größe des Sprach-Korpus
- Größe des Vokabulars
- Art der Trainingsdaten für das Grundmodell
- Lizenzierung
- Geschätzte Kosten für das Finetuning

➢ gpt2 deutsch und gpt2 englisch

parson

Daten aufbereiten

Daten aufbereiten

- XML war geplant, es wurde dann doch PDF
  - Vorverarbeitung aufwändiger
  - Nicht verwertbare Teile wie Kopfzeile, Fußzeile, Weißraum entfernen
- Dokumente werden in Tokens aufgeteilt (tokenizing)
  - Durchflussrichtung > Durch / fluss / richtung
  - 54% der Tokens sind 1 Wort, bei 26% sind 2 Token 1 Wort, bei 13% sind 3 Tokens 1 Wort
- Resultierende Datenmenge wird geteilt
  - 90% geht in das Finetuning, davon 10% zur Validierung während des Finetunings
  - 10% werden für die Bewertung der Vorhersagen zurückgestellt

Endress+Hauser

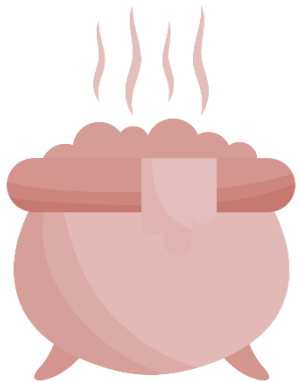
parson

Finetuning

- Hosting: Lokal bei LangTec und AWS
- Tools: Open Source Frameworks, z.B. PyTorch
- Prozess
  - Finetuning: Daten lesen und Gewichtungen anpassen
  - Validation: Vorhersageaufgabe und Feedback
  - Auswertung: Vorhersageaufgaben ohne Feedback

LANG.TEC

SEMANTIC TEXT PROCESSING

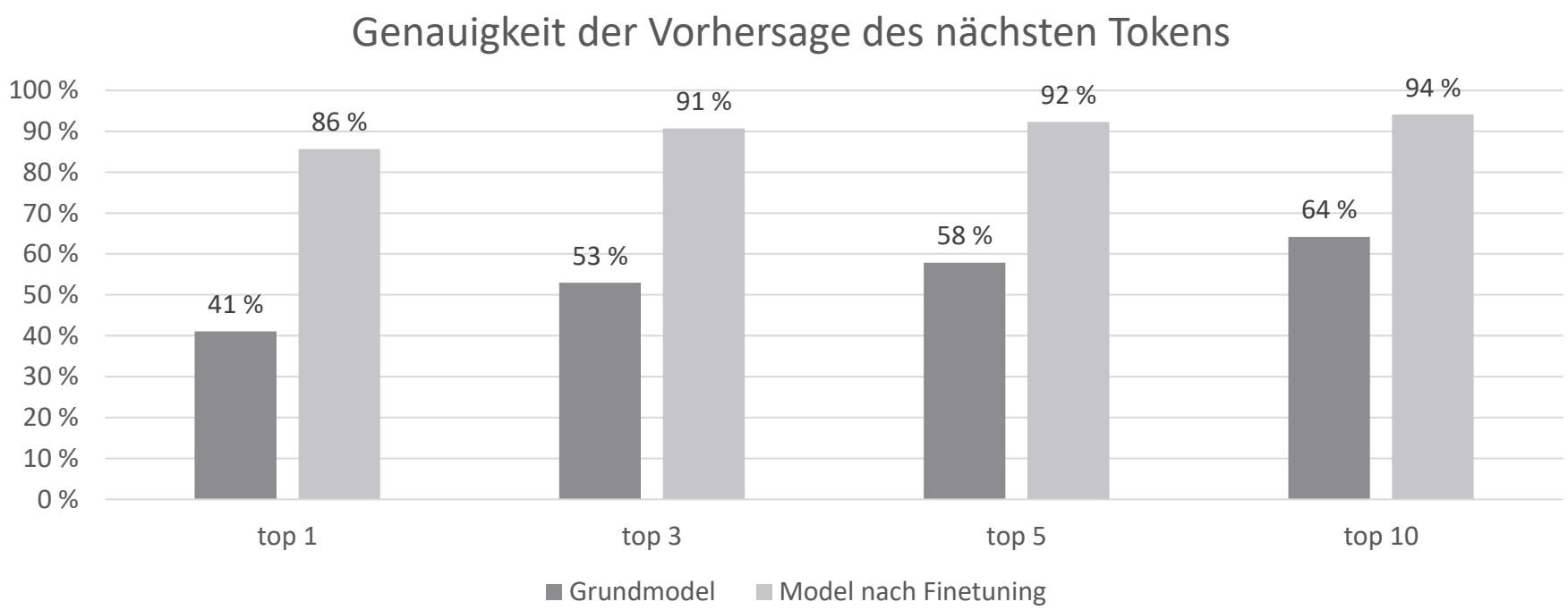


© stockgiu, 123rf.com

parson

Auswertung – Deutsch

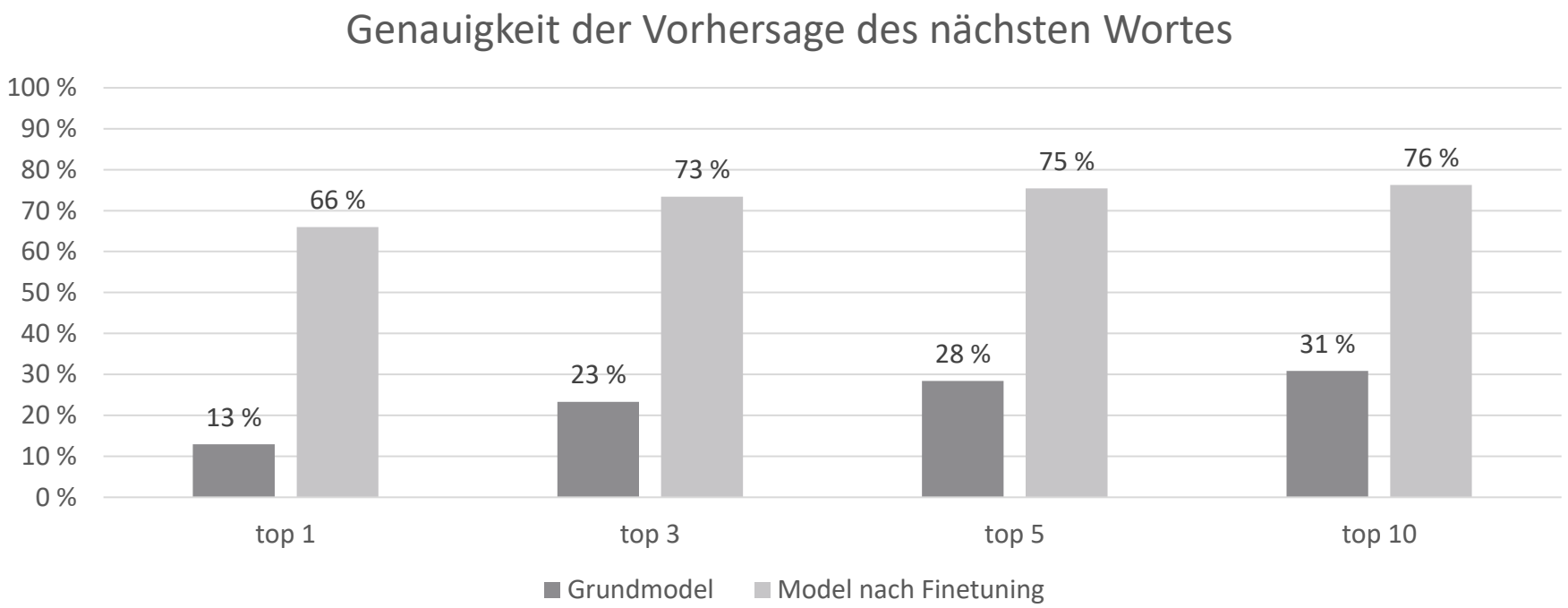
Auswertung



parson

Auswertung – Deutsch

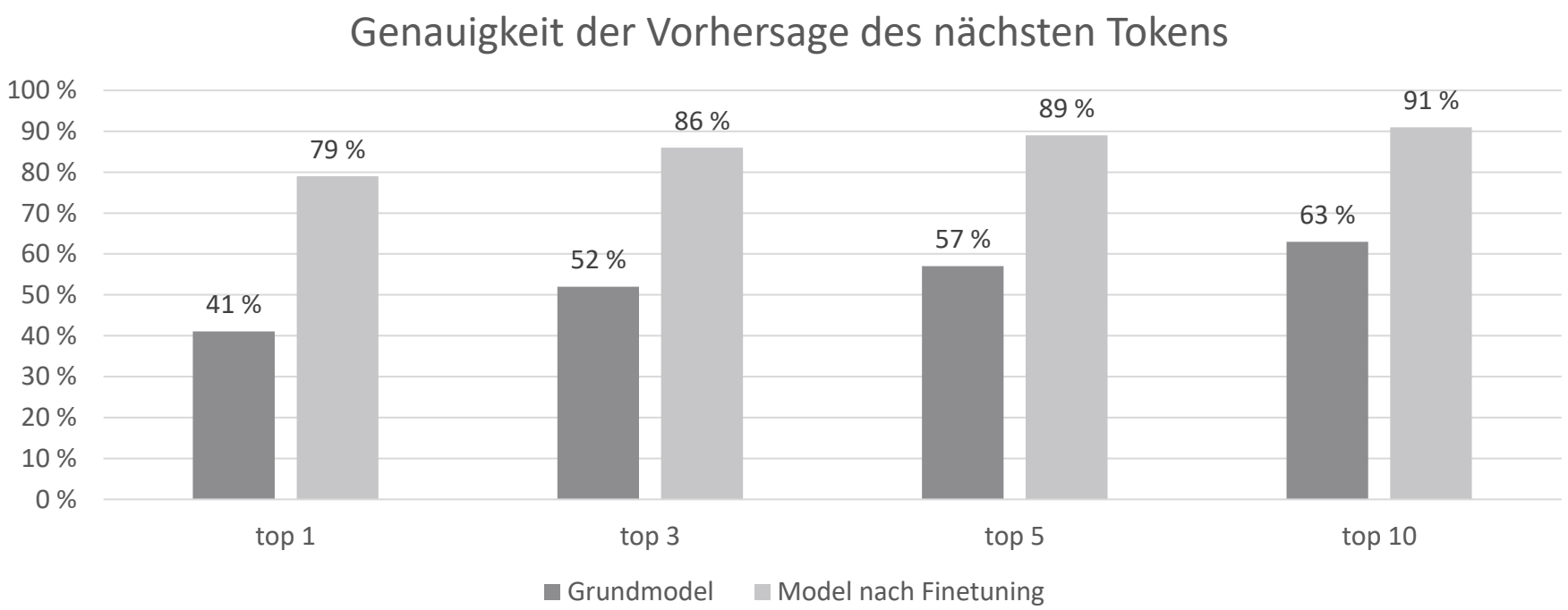
Auswertung



parson

Auswertung - Englisch

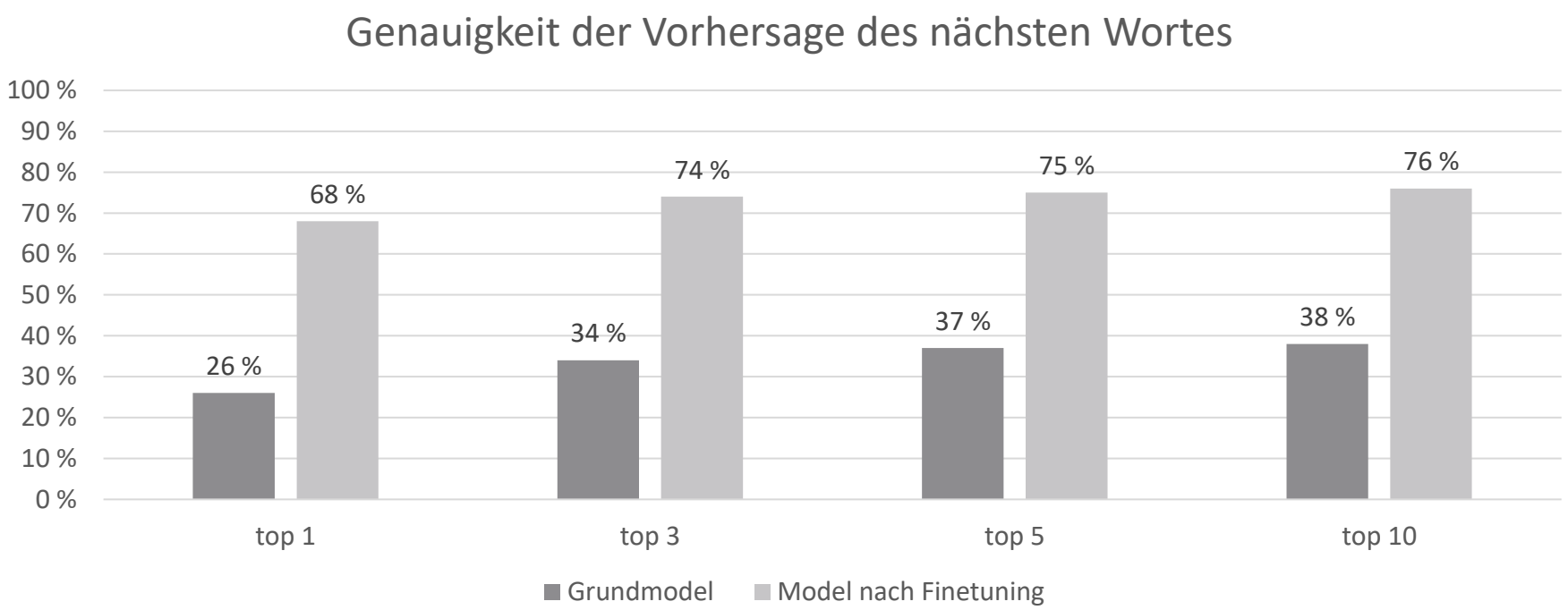
Auswertung



parson

Auswertung - Englisch

Auswertung



parson

Ergebnisse - Grundmodell

Auswertung

Ergebnisse - Finetuning

Auswertung

parson

Integration in Autorenumgebung

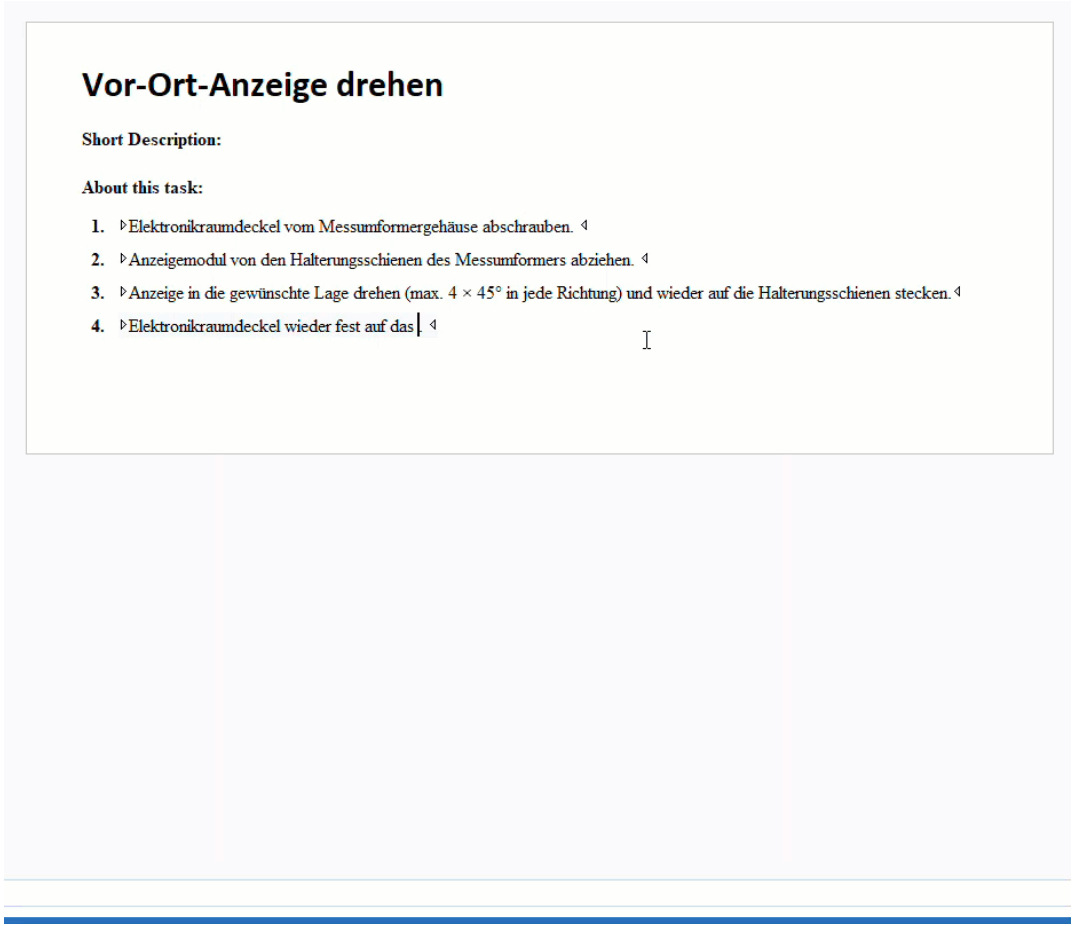
Plugin für Editor entwickeln

- AI Positron Assistant ist eingebettet in Oxygen Author und Editor
- Arbeitet standardmäßig mit OpenAI-Schnittstelle zu ChatGPT
- Wird in Zukunft auch kundenspezifische Sprachmodelle unterstützen
- SyncRO Soft hat vorab Plugin für einfache Anbindung des getunten Modells entwickelt, das Redakteur:in nächste Wörter vorschlägt

syncro soft

<oXygen/> XML Editor

parson

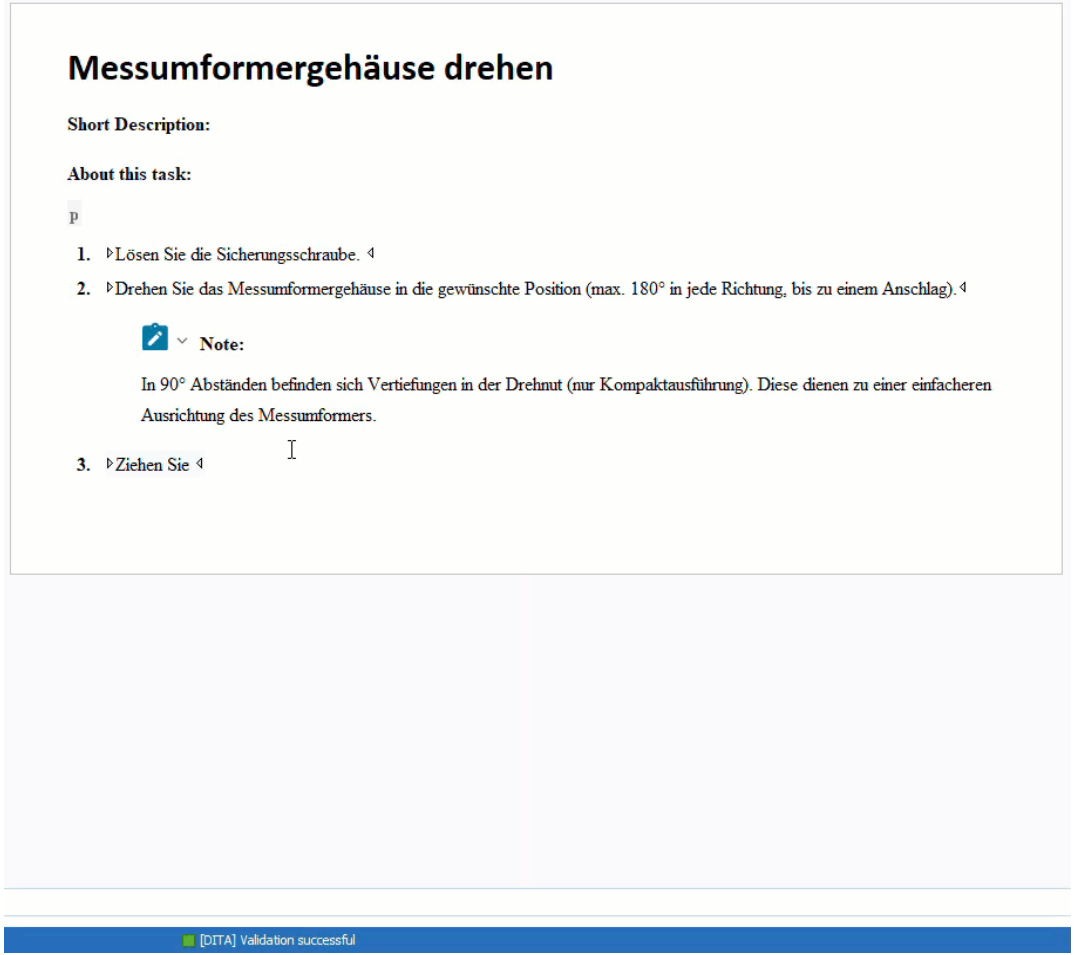


Plugin für Editor entwickeln

syncro soft

<oXygen/> XML Editor

parson



Plugin für Editor entwickeln

syncro soft

<oXygen/> XML Editor

parson

Ausblick auf den produktiven Einsatz in der Technischen Redaktion

parson

Unterschiede Finetuning vs. Prompt-Engineering

Prompt Engineering

- ⊕ Leistungsstarke Sprachmodelle über API nutzbar, z.B. ChatGPT4
- ⊕ Stetige Weiterentwicklung des Modells
- ⊕ Sofort nutzbar, performant

- ⊖ Qualität der Antwort stark vom Prompt abhängig
- ⊖ Prompt abhängig von Autorenumgebung
- ⊖ Keine Anpassung an Unternehmenssprache

Finetuning

- ⊕ Bessere Ergebnisse möglich
- ⊕ Anpassbar an Unternehmenssprache
- ⊕ Open-source Modelle und Frameworks nutzbar
- ⊕ Kleinere Modelle nutzbar

- ⊖ Aufwand für Trainingsdaten und Finetuning
- ⊖ Overfitting möglich
- ⊖ Bei Änderung der Unternehmenssprache neues Finetuning notwendig

parson

Entscheidungshilfe

- Welchen Use Case habe ich?
- Welche Daten habe ich?
  - Wie schützenswert sind die Daten?
  - In welcher Form liegen die Daten vor?
- Welches Budget habe ich?
- Kann ich auf das notwendige Know-how zugreifen?

parson
we create knowledge

www.parson-europe.com

ulrike.parson@parson-europe.com
helle.hannken@parson-europe.com

Danke für Ihre Aufmerksamkeit | Thank you for your attention

Reinbeker Redder 94
21031 Hamburg, Germany

+49 (0)40-7200-500-30
contact@parson-europe.com
parson-europe.com

Newsletter