Guidelines Matching

Qinglin Chen, Hamza Yazbeck, Chenrui Fan

Presenter name

www.islide.cc

What is Guidelines Matching?

Guidelines Matching refers to the process of identifying and aligning relevant guidelines with specific text fragments from a document. Traditionally, this involves manually reviewing the content to determine which guidelines apply to a particular requirement or statement, often requiring significant time and domain expertise.

Why We Need Automatic Approach?

- Ensuring compliance with sustainability guidelines is critical for organizations to meet regulatory and ethical standards.

- Matching text fragments (e.g., from Call for Tenders) with relevant guidelines is essential for automating compliance verification.

- Manual processing of large volumes of text is labor-intensive, highlighting the need for automated systems.

Dataset Overview

Handlungsfeld-ID	Handlungsfeld	Kriterium-ID	Kategorie Kriterium	Ausschreibungskriterium	Ambitionsniveau: Basis	Ambitionsniveau: Gute Praxis	Ambitionsniveau: Vorbild	Quelle
HF-1	Proteine variieren und tierische Produkte reduzieren. Vermehrt vegetarische Gerichte einplanen.	K-1	TS	Das Angebot umfasst eine attraktive Auswahl zur Steigerung des Konsums von Obst, Gemüse, Hülsenfrüchten, Getreide, Nüssen, Kernen/Samen und Pflanzenölen entsprechend der von der schweizerischen Lebensmittelpyramide empfohlenen Nahrungsaufnahme.	Gleich für alle drei Niveaus	Gleich für alle drei Niveaus	Gleich für alle drei Niveaus	Q-1
		K-2	**TS**	**Mindestens ein vegetarisches oder veganes Gericht pro Mahlzeit wird angeboten.**	**Gleich für alle drei Niveaus**	**Gleich für alle drei Niveaus**	**Gleich für alle drei Niveaus**	**Q-1**
		K-3	TS	Mindestens 50% des Tagesangebots sind vegetarisch (von der Gesamtzahl der salzigen Gerichte).	Gleich für alle drei Niveaus	Gleich für alle drei Niveaus	Gleich für alle drei Niveaus	Q-1
		K-4	ZK	Mindestens 50% des Tagesangebots sind vegetarisch (von der Gesamtzahl der salzigen Gerichte).	≥ 60%	≥ 70%	≥ 80%	Q-1
		K-5	TS	Mindestens 1 Tag in der Woche ist 100% vegetarisch (gesamtes Tagesangebot)	Gleich für alle drei Niveaus	Gleich für alle drei Niveaus	Gleich für alle drei Niveaus	Q-1
		K-6	ZK	Mindestens 1 Ta				
				120g				
				120g				
				4 Ge				
		K-10	ZK	Höchstens 4 Ge				
		K-11	TS	100% der vegeta eine leicht erkennbare und				
HF-2	Lebensmittelabfälle vermeiden. Möglichst wenige Lebensmittelabfälle.	K-12	TS	Während minde Lebensmittelabf und identifiziert vermeidbar/unv Auftraggeber zu eingekauften M Resultaten werd				
		K-13	ZK	Während minde Lebensmittelabf und identifiziert vermeidbar/unv Auftraggeber zu eingekauften M Resultaten werden klare Reduktionsziele gesetzt und im Folgejahr überprüft.				

Semantic Similarity: 100%

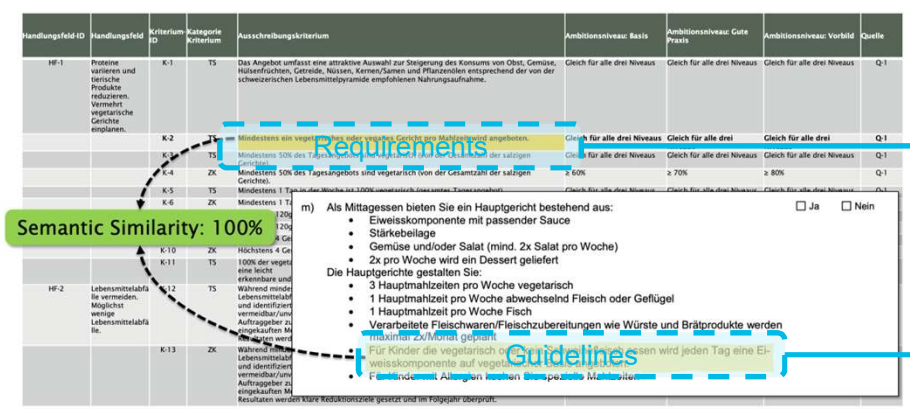
m) Als Mittagessen bieten Sie ein Hauptgericht bestehend aus: ☐ Ja ☐ Nein

- Eiweisskomponente mit passender Sauce
- Stärkebeilage
- Gemüse und/oder Salat (mind. 2x Salat pro Woche)
- 2x pro Woche wird ein Dessert geliefert

Die Hauptgerichte gestalten Sie:

- 3 Hauptmahlzeiten pro Woche vegetarisch
- 1 Hauptmahlzeit pro Woche abwechselnd Fleisch oder Geflügel
- 1 Hauptmahlzeit pro Woche Fisch
- Verarbeitete Fleischwaren/Fleischzubereitungen wie Würste und Brätprodukte werden maximal 2x/Monat geplant
- Für Kinder die vegetarisch oder kein Schweinefleisch essen wird jeden Tag eine Eiweisskomponente auf vegetarischer Basis angeboten.
- Für Kinder mit Allergien kochen Sie spezielle Mahlzeiten

Main Idea



Pretrained LLM

embedding ... embedding

Cosine Similarity

embedding ... embedding

Requirement 1	Guideline 2	Guideline 6	...
Requirement 2	Guideline 5	Guideline 1	...
Requirement 3	Guideline 4	Guideline 13	...
...	...	...	...
Requirement n	Guideline 21	Guideline 16	...

Similarity Table	Guideline 1	Guideline 2	...	Guideline n
Requirement 1	0.47576174	0.70468866	...	0.66516501
Requirement 2	0.02453598	0.23914353	...	0.09460131
Requirement 3	0.5914674	0.44559979	...	0.63468422
...	...	...	...	...
Requirement n	0.93803907	0.22364587	...	0.76866786

Baseline

Baseline: GPT as a Matching Oracle
Directly utilizes GPT to match each requirement with potential guidelines by prompting the model with the requirement text and asking it to identify matching guidelines. This serves as a baseline for comparison with embedding-based methods.

Pipeline 1

Pipeline 1: Cosine Similarity
Uses a pre-trained LLM to generate embeddings for both requirements and guidelines. Computes cosine similarity between each requirement and all guidelines, ranking them based on similarity scores. A straightforward and computationally efficient method.

Pipeline 2

Pipeline 2: Cosine Similarity with Cross-Encoder Re-ranking

Starts with cosine similarity to rank guidelines, then refines the top-K candidates (e.g., top-10) using a cross-encoder architecture. This approach encodes a requirement and a guideline jointly for more context-aware scoring.

Pipeline 3

Pipeline 3: Cosine Similarity with GPT Re-ranking. Similar to Pipeline 2, this approach uses cosine similarity to pre-rank guidelines. However, instead of a cross-encoder, we utilize GPT for re-ranking by providing the top-K guidelines as input and asking the model to determine the most relevant guideline(s) for each requirement.

Pipeline 4

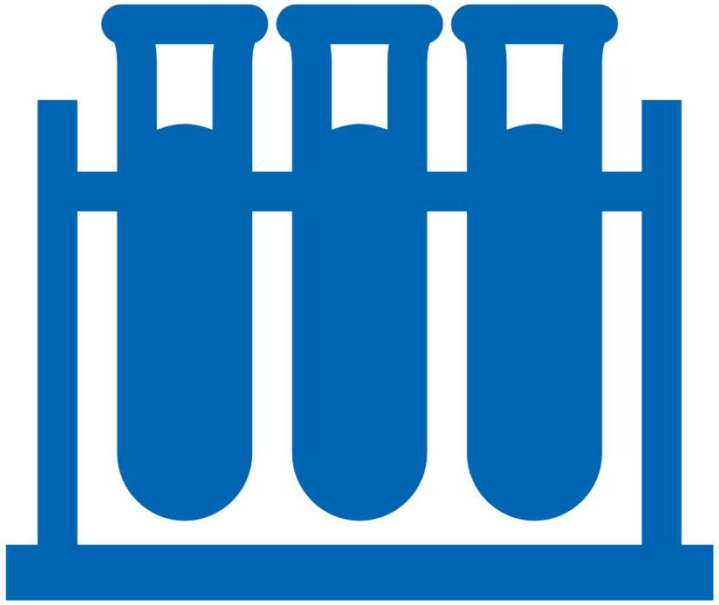
Pipeline 4: Cosine Similarity with Chain of Thought GPT Re-ranking

Builds on Pipeline 3 by incorporating Chain of Thought (CoT) reasoning. GPT is prompted to explain its reasoning step-by-step for each candidate guideline, improving interpretability and accuracy.

Pipeline 5

Pipeline 5: Cosine Similarity with Chain of Thought and Few-shot GPT Re-ranking

Extends Pipeline 4 by adding few-shot examples during GPT prompting. These examples guide the model in its reasoning and matching process, enhancing its ability to re-rank candidates effectively.



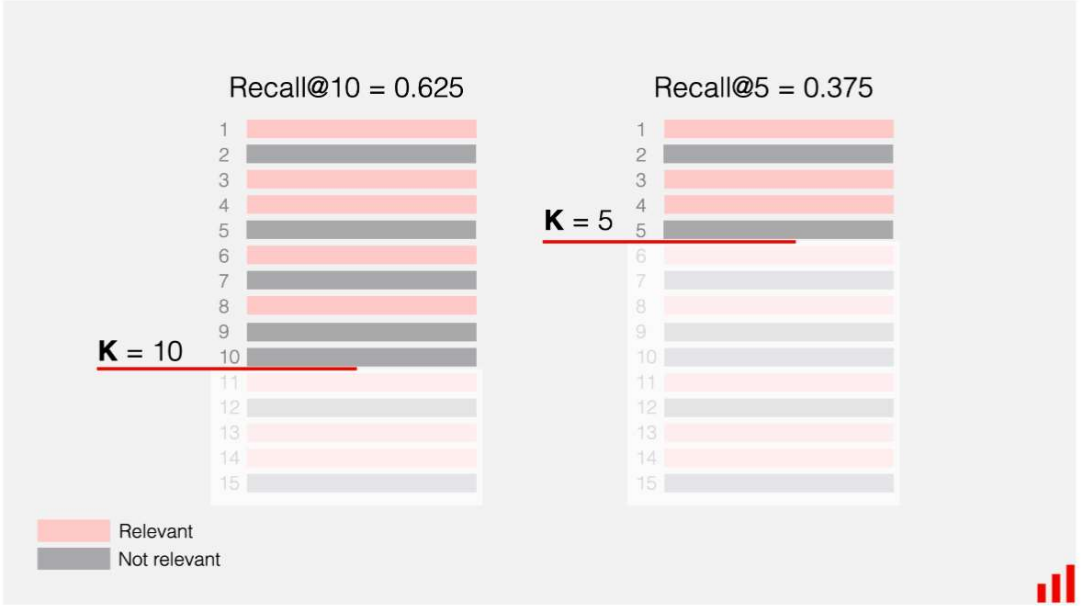
Experiment

Metrics: Recall@K

For requirements that have at least one corresponding guideline, we use Recall@K to measure retrieval performance. Recall@K quantifies the proportion of relevant guidelines successfully retrieved within the top K-ranked results.

$$\text{Recall@K} = \frac{\text{Number of relevant items in the top } K}{\text{Total number of relevant items}}$$

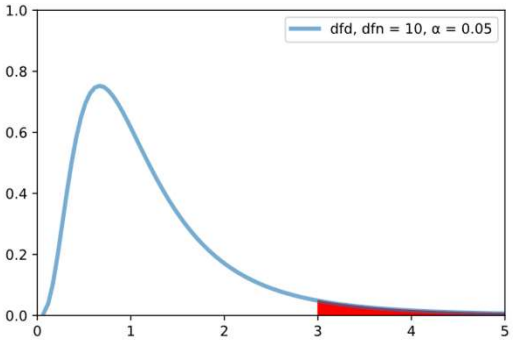
(a) A list of the top 10 recommendations and a total of 8 items in the dataset are actually relevant.



(b) Zoom in on the first 5 suggestions. In this shorter list, we have only 3 relevant recommendations.

Figure 2.1: If the system includes 5 relevant items within the top 10 recommendations, the recall at rank 10 is 62.5%, capturing 5 out of the 8 relevant items in the dataset. Focusing on the top 5 recommendations, the system captures only 3 relevant items. This results in a recall at rank 5 of 37.5%, indicating that the system retrieves less than half of the relevant items in this shorter list.

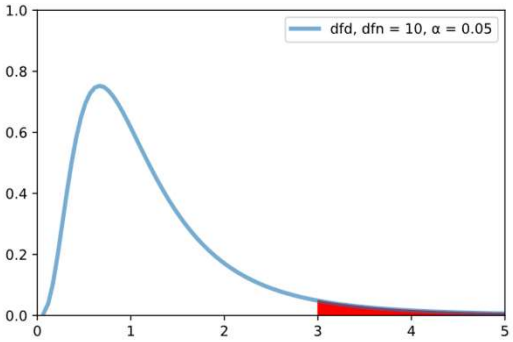
Metrics: F-test



For requirements that lack corresponding guidelines (i.e., without GT), we use an F-test to evaluate the model's ability to identify such cases. The F-test examines the null hypothesis that no guidelines are similar to the given requirement.

$$F_{\text{Stichprobe}} = \frac{S_2^2}{S_1^2} = \frac{\frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(X_{2,i} - \overline{X}_2)^2}{\frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(X_{1,i} - \overline{X}_1)^2}.$$

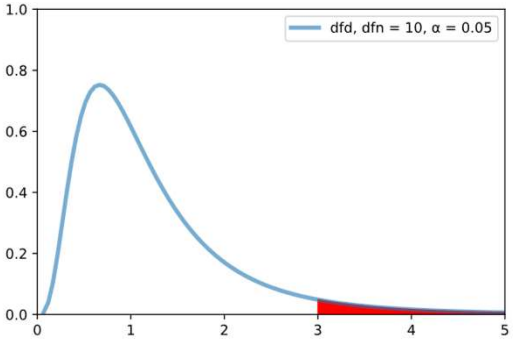
Metrics: F-test



The evaluation involves the following steps:

- Group requirements into two categories: those with GT and those without GT.
- Calculate within-group variances (similarity scores among requirements in the same group) and between-group variances (similarity scores across the two groups).

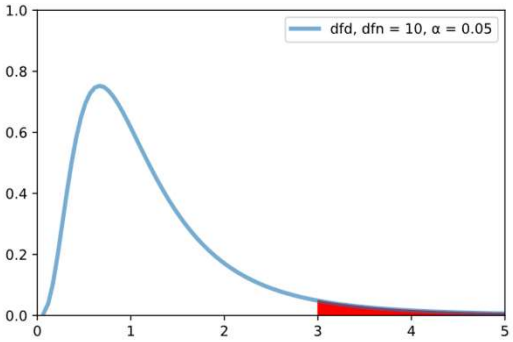
Metrics: F-test



The evaluation involves the following steps:

- Compute a p-value based on the variance ratios. If the p-value exceeds 0.05, the null hypothesis is accepted, indicating no significant similarity between the requirement and any guideline.
- Otherwise, the null hypothesis is rejected, suggesting that at least one guideline is similar to the requirement.

Metrics: F-test



By leveraging prior knowledge of which requirements have GT, this process is framed as a binary classification problem. The classification accuracy reflects the model's ability to identify requirements with no matching guidelines.

Results

Table 3.1: Performance Comparison of Baseline and Pipelines. Metrics include Recall@10, F Accuracy (Unrelated Acc), and Running Time (seconds).

Method	Recall@10	F Accuracy	Running Time (seconds)
Baseline: GPT as a Matching Oracle	0.046±0.023	0.511±0.021	1570.129±1096.301
Pipeline 1: Cosine Similarity	**0.527±0.000**	0.550±0.000	**1.738±0.383**
Pipeline 2: + Cross-Encoder Re-ranking	0.083±0.000	0.383±0.000	5.707±0.513
Pipeline 3: + GPT Re-ranking	0.340±0.069	**0.572±0.016**	272.409±0.575
Pipeline 4: + Chain of Thought	0.202±0.153	0.567±0.036	542.385±0.357
Pipeline 5: + Few-shot GPT Re-ranking	0.098±0.060	0.539±0.028	2066.009±15.907

Results

"TS Das Angebot umfasst eine attraktive Auswahl zur Steigerung des Konsums von Obst, Gemüse, Hülsenfrüchten, Getreide, Nüssen, Kernen/Samen und Pflanzenölen entsprechend der von der schweizerischen Lebensmittelpyramide empfohlenen Nahrungsaufnahme.“

1. "To match the guideline regarding the attractive selection of food items to increase the consumption of fruits, vegetables, legumes, grains, nuts, seeds, and plant oils according to the Swiss food pyramid, we can identify the most relevant requirements from the provided list. Here’s a ranking based on relevance:"
2. ""Die Qualität der Ernährung an den Bieler Tagesschulen und Kitas neu entsprechend dem Reglement «Gesunde Ernährung» sichergestellt werden."
3. This requirement directly relates to ensuring the quality of nutrition, which aligns with the guideline's focus on increasing the consumption of recommended food groups.“
4. ""Sie verwenden Öle, Gewürze, Streusalze, Zucker, Backwaren, wie auch weitere Basisprodukte und Nahrungsmittel für die Zwischenmahlzeiten zu 50% (an Kosten gemessen) in der Qualität «Bio» und (bei Produkten aus der 3. Welt) mit dem Gütesiegel «Max Havelaar».""**"

Thank you

Presenter name

www.islide.cc