LabelSense: Evaluation & Model Summary

LabelSense: Model & Approach Decision Matrix

Task / Component	Preferred Approach	Reason for Using	Backup / Alternative
OCR	EasyOCR	Open-source, multilingual, fast	Google OCR, Tesseract
Ingredient NER	BERT-NER	Structured, high F1, adaptable	GPT-4, dictionary lookup
Dietary Classification	RoBERTa	High accuracy, multi-label	GPT-4, LightGBM
Allergen Detection	Hybrid: Rule + GPT-4	Rules for explicit, LLM for inferred	Rules only, ML classifier
Ingredient Explanation	GPT-4 + Caching	Clear, human-like, simple	Wikipedia/API, BART, T5
Personalization	Rules + GPT	Deterministic + flexible fallback	DistilBERT, fine-tuned model
Evaluation & UX	Quantitative + Survey	Captures performance + user feedback	NA

LabelSense: Enhanced Evaluation Metrics Matrix

Component	Primary Metric(s)	Why	Secondary	Human Eval?
OCR	CER, WER	Text accuracy vs ground truth	BLEU, Levenshtein	No
NER	F1-Score	Balance precision/recall	Precision, Recall	No
Diet Class.	Multi-label Acc., F1	Handles multi-tag, imbalance	Hamming Loss, ROC-AUC	Trust Score (1-5)
Allergen	Recall	Safety-critical, catch all	F1, Precision	Trust Score (1-5)
Explanation	User Score (1-5)	Clarity and comprehension	BLEU, ROUGE	Pairwise Ranking
Personalization	Recall (Profile-based)	Matches user needs	False Neg. Rate	Scenario Test
System UX	SUS, Task Time	Ease, trust, satisfaction	Qual. Feedback	SUS Survey

Human Feedback Analysis Methods

Analysis Type	Action	Tool / Output	Purpose
Quantitative	Calculate mean/std	Pandas, Excel	Summarize ratings (trust, SUS)
Quantitative	Visualize scores	Matplotlib, seaborn	Compare models or groups
Quantitative	Correlate scores	Spearman, Pearson	Link metrics to system performance
Qualitative	Code open feedback	Manual, Taguette	Find recurring themes
Qualitative	Highlight quotes	Markdown/LaTeX	Support qualitative claims
Qualitative	Word cloud	Python, online tools	Visualize frequent terms