LabelSense: Evaluation & Model Summary

Model & Approach Decision Matrix

Component	*Approach*	*Reason*	*Alternative Approach*
OCR	EasyOCR	Open-source, multilingual, fast	Google OCR
Ingredient NER	BERT-NER	Structured, high F1, adaptable	GPT-4, dictionary lookup
Dietary Classification	RoBERTa-base	High accuracy, multi-label	GPT-4
Allergen Detection	Hybrid: Rule + GPT-4	Rules for explicit, LLM for inferred	N/A
Ingredient Explanation	GPT-4 + Caching	Clear, human-like, simple	Wikipedia/API
Personalization Logic	Rules + GPT	Deterministic + flexible fallback	DistilBERT
Evaluation	Quantitative + Survey	Captures performance + user feedback	N/A

Evaluation Metrics Matrix

Component	*Primary Metric(s)*	*Reason for Metric*	*Secondary Metric(s)*
OCR	Character Error Rate (CER), Word Error Rate (WER)	Measures how accurately OCR matches true ingredient text	Levenshtein distance, BLEU
Ingredient NER	F1-Score (overall + per-class)	Evaluates balance between precision and recall across ingredient types	Precision, Recall
Dietary Classification	Multi-label Accuracy, Weighted F1	Supports multi-tag dietary detection; handles class imbalance	Hamming Loss, ROC-AUC
Allergen Detection	Recall (primary), Per-allergen Recall	Prioritizes catching all allergens—critical for user safety	Precision, False Positive Rate
Ingredient Explanation	User Comprehension Score (1–5), Explanation Accuracy	Measures how clearly explanations are understood	BLEU/ROUGE (optional for validation)
Personalization Logic	Profile-based Recall, Manual validation	Validates filtering based on individual user needs	False Negative Rate per profile
Application Evaluation	SUS, Task Completion Time, Trust Rating	Captures user satisfaction, system trust, and ease of use	Qualitative feedback

Human Feedback

Component	*Human Survey*
OCR	✘
Ingredient NER	✘
Dietary Classification	✔ Trust rating (1–5): "Do you trust the dietary classification?"
Allergen Detection	✔ Trust score (1–5): "Would you rely on this allergen alert?"
Ingredient Explanation	✔ Pairwise ranking: GPT-4 vs. baseline explanations
Personalization Logic	✔ Scenario-based test: "Was the ingredient correctly flagged for you?"
Application Evaluation	✔ SUS 10-question survey + open feedback

Human Feedback Analysis

Analysis Type	*Action*	*Tool / Output*	*Purpose*
Quantitative	• Calculate mean/std • Visualize scores • Correlate scores	Pandas, Excel Matplotlib, Seaborn Spearman, Pearson	• Summarize ratings (trust, SUS) • Compare models or user groups • Link metrics to system performance