Structured Extraction from Catering Offer Emails

Automating the extraction of key offer details from unstructured email text using NLP techniques

The Challenge

Manual Offer Comparison Creates Bottlenecks

Current Reality

- Event organizers send RFPs to multiple catering suppliers
- Offers arrive as free form emails with varying formats
- Manual extraction into spreadsheets is time consuming and error-prone
- Critical details may be missed or misinterpreted

Key Challenges

- **Unstructured text:** Different writing styles across suppliers
- **Format variability:** Prices, dates, policies appear in different places
- **Missing information:** Fields may be implicit or omitted entirely
- **Inconsistent units:** "around 200 per person", "+VAT", ranges

Project Overview: Email-to-JSON Extraction

Input

Raw catering offer email (subject + body) in natural English

Processing

NLP model extracts structured fields using LLM or fine tuned approach

Output

Standardized JSON with supplier name, pricing, guests, VAT, kosher level, bar, dates, policies

Novel Contribution: We propose a new synthetic dataset and a systematic evaluation

Model Pipeline & Synthetic Data Generation

01

Generate Ground Truth JSON

Sample realistic values: prices, guest counts, VAT status, kosher options, bar inclusion, dates, cancellation policies

02

Create Synthetic Emails

LLM generates business style English emails from JSON templates with natural variation and noise

03

Train & Evaluate Models

Compare zero shot LLM, few shot LLM with examples, and fine tuned smaller open-source models

04

Validate Predictions

Compare predicted JSON against ground truth, compute field wise and JSON level accuracy metrics

Zero Shot Baseline

Schema description only

Few Shot Baseline

Schema + example pairs

Fine Tuned Model

Supervised on synthetic dataset

Metrics and KPIs

How would you measure the results?

Extraction Quality

- Field wise performance on key numeric, boolean, and categorical fields
- "All key fields correct" accuracy per email (JSON level accuracy)

Data Generation Quality

- Spot check a sample of generated emails to ensure they are realistic and consistent with their JSON

Comparison Protocol

- Compare the model's predicted JSON (pred_json) with the ground truth JSON (gold_json) for each email
- Compute metrics per field and overall

Boolean / Categorical Fields

- Examples: includes_vat, is_kosher, includes_bar, price_type
- Metric: Accuracy (and optionally F1)

Numeric Fields

- Examples: price_value, min_guests, max_guests
- Metrics: Exact match accuracy; optionally MAE (mean absolute error)

Short Text Fields

- Examples: kosher_supervision, date_available
- Metric: Exact match after simple normalization

Key Success Metric: "Key-fields JSON accuracy" - percentage of emails where all core fields (price, VAT, kosher, min guests, bar) are predicted correctly. Ground truth comes directly from the original JSON used to generate each synthetic email.