

Supplementary Material S1

AI-based structured extraction pipeline: prompts, JSON schema, and a representative output

This supplement provides the complete specification of the semi-automated extraction pipeline used to derive structured content variables (database used, machine-learning methods, other statistical methods, study class, clinical outcome, patient count and AUROC) from publication titles and abstracts. All calls used the DeepSeek-V2 model (DeepSeek AI) via the DeepSeek API. For each publication the model received the title and abstract together with the JSON schema below and returned a single JSON object. A second, independent call was used for country assignment. The material is provided here rather than in the main text because of its length.

1. Primary extraction system prompt

```
SYSTEM_PROMPT = (  
    "Review given article. "  
    "Answer in json [field_name:answer, ...] "  
    "the most applicable answer option for query. "  
    "Strictly answer with the possible answer, "  
    "but if it is false or null skip field. "  
    "if answer options are in parentheses answer strictly. "  
    "For 'db_used', explicitly mark each database as true or false based on the text."  
)
```

2. Extraction JSON schema

Each leaf carries a natural-language `llm_query`. Boolean fields (`db_used`, `machine_learning_method`, `other_method`) are evaluated independently, so a single publication can be positive for several labels (multi-label design).

```
SCHEMA = {  
    "db_used": {  
        "mimic4": {  
            "llm_options": [True, False],  
            "llm_query": "used data from MIMIC IV / Medical Information Mart for Intensive Care IV"},  
        "mimic3": {  
            "llm_options": [True, False],  
            "llm_query": "used data from MIMIC III / Medical Information Mart for Intensive Care III"},  
        "sicdb": {  
            "llm_options": [True, False],  
            "llm_query": "used data from SICdb / Salzburg Intensive Care Database"},  
        "eicu": {  
            "llm_options": [True, False],  
            "llm_query": "used data from the eICU dataset"},  
        "hirid": {  
            "llm_options": [True, False],  
            "llm_query": "used data from HiRID / High-Time-Resolution ICU Dataset"},  
        "amsterdam": {  
            "llm_options": [True, False],  
            "llm_query": "used data from the AmsterdamUMCdb dataset"}  
    },  
    "amount_of_patients_included": {},  
    "study_class": {  
        "retrospective_observation": {"llm_query": "study is a retrospective observation"},  
        "predictive": {"llm_query": "study does predictive modeling"},  
        "decision_support": {"llm_query": "study does decision support modeling"},  
        "data_science": {"llm_query": "study is about the database itself"},  
        "other": {"llm_query": "other options do not apply"}  
    },  
    "machine_learning_method": {  
        "logistic_regression": {"llm_options": [True, False], "llm_query": "logistic regression used"},  
        "linear_regression": {"llm_options": [True, False], "llm_query": "linear regression used"},  
        "lasso_regression": {"llm_options": [True, False], "llm_query": "lasso regression used"},  
        "knn": {"llm_options": [True, False], "llm_query": "k-nearest neighbors used"},  
        "decision_tree": {"llm_options": [True, False], "llm_query": "decision tree used"},  
        "random_forest": {"llm_options": [True, False], "llm_query": "random forest used"},  
        "support_vector_machine": {"llm_options": [True, False], "llm_query": "support vector machine used"},  
        "gradient_boosting": {"llm_options": [True, False], "llm_query": "gradient or similar boosting used"},  
        "neural_network": {"llm_options": [True, False], "llm_query": "neural network used"},  
        "reinforcement_learning": {"llm_options": [True, False], "llm_query": "reinforcement learning used"}  
    },  
    "other_method": {  
        "cox_proportional_hazard_regression": {"llm_options": [True, False], "llm_query": "cox proportional hazard regres"},  
        "bayes": {"llm_options": [True, False], "llm_query": "bayes statistics used"},  
        "propensity_score": {"llm_options": [True, False], "llm_query": "propensity score used"},  
        "nomogram": {"llm_options": [True, False], "llm_query": "nomogram was created"}  
    },  
    "outcome": {
```

```
"AKI": {"llm_query": "Acute Kidney Injury outcome studied"},
"mortality": {"llm_query": "Mortality outcome studied"},
"icu_length_of_stay": {"llm_query": "ICU Length of Stay outcome studied"},
"other": {"llm_query": "Other specific outcomes studied"}
},
"AUROC": {"llm_query": "Area Under the Receiver Operating Characteristic curve reported"}
}
```

3. Secondary prompt (country assignment)

Country was assigned with a separate call, concatenating all <affiliation> tags of a paper into one text block.

```
secondPrompt = (  
  'Extract the main Country of the researchers or the study location based on the  
  provided text (Title, Abstract, or Affiliations). Respond ONLY in valid JSON  
  format with the key country. If the country cannot be reasonably determined,  
  output null for country. Do not guess without evidence.'  
)
```

4. Representative model input and output

The example below illustrates a typical multi-label case: a retrospective MIMIC-IV mortality-prediction study comparing several models. The abstract text shown is a paraphrased illustration; the JSON reflects the schema above. Only fields evaluated as true/non-null are emitted, per the system prompt.

Input (title + abstract, abbreviated):

```
Title: "Machine learning models for predicting in-hospital mortality in  
       sepsis patients: a retrospective analysis of the MIMIC-IV database"  
Abstract: "... 8,432 septic patients from MIMIC-IV were included. Logistic  
          regression, random forest and XGBoost were compared. The XGBoost model  
          achieved an AUROC of 0.87 ... "
```

Output (JSON):

```
{  
  "db_used": {"mimic4": true},  
  "amount_of_patients_included": 8432,  
  "study_class": ["retrospective_observation", "predictive"],  
  "machine_learning_method": {  
    "logistic_regression": true,  
    "random_forest": true,  
    "gradient_boosting": true  
  },  
  "outcome": ["mortality"],  
  "AUROC": "0.87"  
}
```

5. Post-processing

AUROC strings were scanned left-to-right and the first decimal or valid percentage in the range 0.50-1.00 was retained; values that could not be parsed or fell outside these bounds were set to missing. Patient counts below 10 were treated as extraction artefacts and excluded from cohort-size analyses. Records without a determinable country were labelled "Unknown". Field-level extraction reliability (precision, recall, F1) against a manually coded reference standard is reported in Table 5 of the main manuscript.