

Supplementary Material 1

GeneResolver: A Traceable Hybrid Multi-Agent Pipeline for Etiology-Aware Gene Prioritization

1 De-identification test set

The dataset was generated with Claude Sonnet 5 and contains 100 synthetic notes. Annotated tokens were verified as exact case-insensitive substrings.

You are generating a synthetic clinical de-identification test set. Produce ONE realistic but entirely FICTIONAL clinical referral or consultation note per item, written in the style of a rare-disease genetics referral. The note may include clinically relevant findings such as seizures, developmental delay, dysmorphic features, metabolic abnormalities, and skeletal or connective-tissue findings.

Every patient, relative, clinician, facility, location, and identifying number must be invented. Do not use the name of a real person, public figure, hospital, clinic, or other healthcare institution. Do not reuse a real medical record number or any other real identifying information.

Embed a natural mixture of information that a clinical de-identification system should detect and mask. The possible information fields include:

- patient, relative, or clinician names;
- street addresses, cities, postal codes, and other detailed geographic locations;
- dates linked to an individual, including dates of birth, admission dates, discharge dates, and consultation dates;
- telephone and fax numbers;
- email addresses;
- medical record numbers;
- health-plan or insurance identifiers;
- account numbers;
- professional licence or certificate numbers;
- vehicle identifiers;
- medical-device identifiers or serial numbers;
- URLs;
- IP addresses;
- biometric identifiers represented in text;

- other unique identifying numbers or codes;
- ages above 89 years.

Vary the information fields included in each note. Do not include every field in every note. Include a minority of notes, approximately one in six, with no identifying information so that the dataset also contains true-negative examples.

For EACH note, output the exact text substrings that should be detected and masked. Copy every annotated substring VERBATIM from the note, preserving the same characters and capitalization. These annotations will be checked programmatically against the note text. For example, if the note contains "DOB 03/15/1985", the annotated token must be exactly "03/15/1985", rather than "March 15, 1985" or "1985".

Return one JSON object per line, without Markdown fencing:

```
{"id": "PHI_ONN", "note": "...",  
"expected_pii_tokens": ["...", ...]}
```