

Supplementary Information for Auditing Data Leakage Candidate Coverage and Calibration in Clinical Abbreviation Disambiguation Benchmarks

Tianze Yang^{1*}, Qiqing Li², Jialun Wu³, Xinyu Qiu⁴

^{1*}Independent Researcher, San Jose, 95134, CA, USA.

²Independent Researcher, Mountain View, 94043, CA, USA.

³Independent Researcher, Jersey City, 07306, NJ, USA.

⁴Northeastern University, Boston, 02115, MA, USA.

*Corresponding author(s). E-mail(s): tianzeyy@outlook.com;
Contributing authors: qiqingli3006@outlook.com; jialunw@outlook.com;
qiu.xiny@northeastern.edu;

Minimum reporting checklist

Table [S1](#) summarizes the minimum information a reader needs in order to reproduce and interpret an abbreviation benchmark audited with the protocol described in the main text. The checklist is intended for any corpus with an abbreviation, sense and context representation, not only CASI, and follows the same construction stages described in Section 3.2 of the main text.

Table S1 Minimum reporting checklist for leakage aware and coverage auditable clinical abbreviation benchmarks. The final column gives the corresponding section of the main text.

Construction stage	Item to report	Where reported in main text
Filtering Duplicate isolation	Raw row count and the label filter applied Surplus duplicate rows, conflicting groups, resulting unique context count, and sensitivity to the identity definition	Sec. 3.2, Sec. 4.1, Sec. 5.1 Sec. 3.2, Sec. 5.1, Sec. 5.7
Target definition	Class definition used for the prediction target	Sec. 3.2
Candidate coverage routing	Minimum support threshold, core and rare class counts, and class and instance fractions	Sec. 3.2, Sec. 5.1
Partitioning	Split seed and manifest, plus accuracy across several adjacent seeds	Sec. 3.2, Data availability, Sec. 5.6
Reference selector	Selector, hyperparameters and library versions	Sec. 3.2, Sec. 4.1, Table 5, Appendix A
Candidate ranking	Candidate recall at rank, mean reciprocal rank and pool sizes	Sec. 5.6
Calibration and abstention	Calibration before and after scaling, the risk and coverage curve and confidence on the coverage audit	Sec. 3.2, Sec. 5.3, Sec. 5.6, Fig. 3
Resource measurement	Training time, model size, throughput and latency, with hardware and one size unit family	Sec. 3.2, Table 5