CAN INDIVIDUAL PHYSICIAN HISTORY IMPROVE DIALOGUE-ACTION PREDICTION? A PERSONALIZATION BENCHMARK WITH FOUR PROFILE SOURCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Profile-based personalization uses individual statistical priors, history retrieval, dense profile representations, and user-specific parameter adaptation. These methods use history differently, but they are usually evaluated in the same way: predictions made with the target individual's profile are compared with predictions made without a profile. An improvement in this comparison does not by itself show that the model uses history specific to that individual. Relevant histories from similar individuals may already help, while a profile from another individual may harm prediction through mismatch. We study this issue in physician dialogue-action prediction. For each target physician reply, we hold the current dialogue fixed and vary only the profile source, comparing no profile, a department-pooled profile, another physician's profile, and the target physician's profile. The pooled profile excludes the target physician, and the other physician comes from the same department and has a similar number of profile consultations. We construct a benchmark with 244 evaluation physicians and 164,702 physician-reply events in the test set, labeled with 11 functional dialogue actions. We evaluate seven representative methods spanning statistical priors, history retrieval, dense profile representations, and user-specific parameter adaptation. Some gains remain after the target physician is excluded, while some Target-versus-Other differences arise because the other physician's profile increases the loss. Same-department history retains part of the profile gain for several methods, whereas the additional improvement from the target physician's history varies by method. Comparing all four sources shows whether an apparent gain remains with same-department history, depends on the target physician's records, or is enlarged by a mismatched profile.

1 INTRODUCTION

Personalization from historical profiles has become an important direction in language-model research. Existing methods represent and use individual history in several ways. Some directly model long-term preferences or behavioral frequencies; others retrieve past interactions relevant to the current input, compress a complete history into a dense profile representation, or encode individual information in user-specific parameter modules (Salemi et al., 2024; Tan et al., 2024; Liu et al., 2025; Park et al., 2026). Despite substantial differences in architecture and history use, these methods are usually evaluated by comparing predictions made with the target profile against predictions made with no profile.

This comparison answers whether history helps, but not whether the improvement depends on history from the target individual. If histories from similar individuals are already useful, a Target-versus-No gain does not by itself show that target-specific history produced the improvement. Conversely, if a coherent profile from another individual is mismatched to the target, its poorer performance can enlarge the Target-versus-Other difference. With only Target Profile and No Profile, we cannot tell how much of the gain is already available from same-department history or what happens when the model receives another physician's profile.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We address this issue with a four-source evaluation for physician dialogue-action prediction. For the same target physician reply, we hold the dialogue, prediction target, and model settings fixed and change only the available profile: no profile; pooled histories from the same department excluding the target physician; the profile of another physician from the same department with a similar number of profile consultations; or the target physician's own profile. The comparisons show whether same-department history is already useful, whether the target physician's own history adds further benefit, and whether another physician's history interferes with prediction.

We construct the benchmark from public Chinese multi-turn online consultations (Lu et al., 2022) and evaluate it on 244 physicians and 164,702 physician-reply events in the test set, labeled with 11 functional dialogue actions. Seven predictors cover statistical priors, context-based individual residuals, history retrieval, dense profile representations, and user-specific parameter adaptation. Improvement with the target profile is not equivalent to using information specific to that physician. For some methods, same-department histories retain a substantial share of the gain after excluding the target physician. The additional improvement from the target physician's own history varies markedly, and some methods perform worse when given another physician's profile. The pooled condition shows how much of the gain remains without the target physician; the other-physician condition shows whether substituting another physician's history helps or harms prediction.

Our contribution has three parts:

Four profile sources evaluated on the same event. We compare no profile, pooled same-department history, another individual's profile, and the target individual's profile. The comparisons show whether a target-profile gain remains with same-department history and how prediction changes when another individual's history is substituted.

Seven methods tested under the same source substitutions. The predictors span statistical, retrieval-based, dense-representation, and user-specific parameter approaches. Each method is evaluated on the same events under all four sources.

Different explanations for an apparent personalization gain. Some gains are already present with same-department history, while some Target-versus-Other differences are enlarged by poorer performance under another individual's profile.

2 Related Work

2.1 Personalization from Historical Profiles

Personalization from historical interactions has developed along several technical paths. Early dialogue systems represented user differences through explicit personas or interaction histories (Li et al., 2016; Zhang et al., 2018). With large language models, history enters prediction in richer forms. LaMP retrieves records relevant to the current input and supplies them as additional context (Salemi et al., 2024), while PPlug builds a continuous user representation from a broader history and conditions prediction through a soft representation (Liu et al., 2025). Other methods encode individual information in model parameters: OPPU learns parameter-efficient adaptation modules for individual users (Tan et al., 2024), and PRISP constructs user-related parameter representations from a small amount of personal history (Park et al., 2026). Simpler methods based on historical frequencies, individual residuals, or similar-state retrieval can also use repeated behavioral records directly.

These approaches extend the representation of history from discrete records to continuous representations and user-specific parameters, but share the aim of using past behavior to improve current prediction.

2.2 Profile Evaluation and History-Source Controls

Although history enters these models in different ways, personalization gains are usually measured by comparing the target-profile and no-profile conditions (Salemi et al., 2024; Tan et al., 2024; Liu et al., 2025; Park et al., 2026). This comparison shows whether adding target history improves prediction, but not whether the improvement depends on that history belonging to the target individual. Histories from people with similar backgrounds may already contain useful information, while a complete

profile from someone else may affect the model through mismatch. Target-versus-No alone therefore cannot distinguish general historical information, shared background, and the target individual's history.

Previous work has examined profile content through several controls. PersonaLens ablates sources of user context, PERG compares relevant, irrelevant, mixed, and missing preferences, and PENGUIN studies model behavior under different context conditions (Zhao et al., 2025; Okite et al., 2025; Wu et al., 2025). Random histories or profiles have also been used to test whether personalized models depend on user-related information. Lu et al. (2022) use mismatched profile–dialogue pairs as negative examples when learning physician representations.

These studies examine missing profiles, content relevance, or incorrect history. We place no profile, pooled same-department history, another individual's history, and the target individual's history in the same evaluation and compare their effects on the same prediction event.

2.3 Physician Dialogue Actions and Repeated Individual Behavior

Medical dialogue datasets support tasks such as response generation, history taking, and medication recommendation (Zeng et al., 2020; Saley et al., 2024; He et al., 2022; Shi et al., 2024). EHR and policy-learning studies instead examine future clinical events, treatment choices, or observed decision trajectories (Choi et al., 2016; Deuschel et al., 2024; Hüyük et al., 2022; Futoma et al., 2020; Raghu et al., 2017; Komorowski et al., 2018). Health-services research has also documented persistent differences in physician practice, which are intertwined with patient mix, department background, and organizational context (Cutler et al., 2019; Molitor, 2018; Finkelstein et al., 2016; Chan et al., 2022).

Physicians who appear repeatedly in online consultations provide a natural setting for comparing history sources. Each physician has multiple distinct consultations and a defined department, allowing us to construct the target physician's history, pooled same-department history, and another physician's complete history and compare them under the same consultation context. We represent physician replies as functional dialogue actions and predict the physician's next behavior in the current consultation.

3 Benchmark Construction and Evaluation

We examine whether physician history improves dialogue-action prediction and whether that improvement depends on the target physician's own history. Figure 1 summarizes the design.

3.1 Task and Benchmark Overview

We construct a dialogue-action prediction benchmark from physicians with repeated consultations. Each physician reply is a prediction event, and the preceding dialogue is used to predict the functional actions in that reply (Figure 1A). For event i, let x_i denote the visible dialogue, d_i the target physician, and $y_i \in \{0,1\}^{11}$ the multilabel actions in the target reply. Given a history source $P(d_i)$, the model outputs action probabilities:

$$p_i = f(x_i, P(d_i)) \in (0,1)^{11}.$$

For the same prediction event, the dialogue, target, and model settings remain unchanged. We replace only the history supplied to the personalization component: no profile, a department-pooled profile, a matched other physician's profile, or the target physician's profile (Figure 1B). The pooled profile excludes the target physician, while the other profile comes from a physician in the same department with a similar number of profile consultations. We compare seven methods spanning statistical aggregation, retrieval, dense profile representations, and parameter adaptation, using binary cross-entropy as the primary metric (Figure 1C). Comparisons with no profile, the department-pooled profile, and the other physician's profile show the overall change after adding target history, the improvement already available from same-department history, and the role of mismatch in Target-versus-Other differences (Figure 1D).

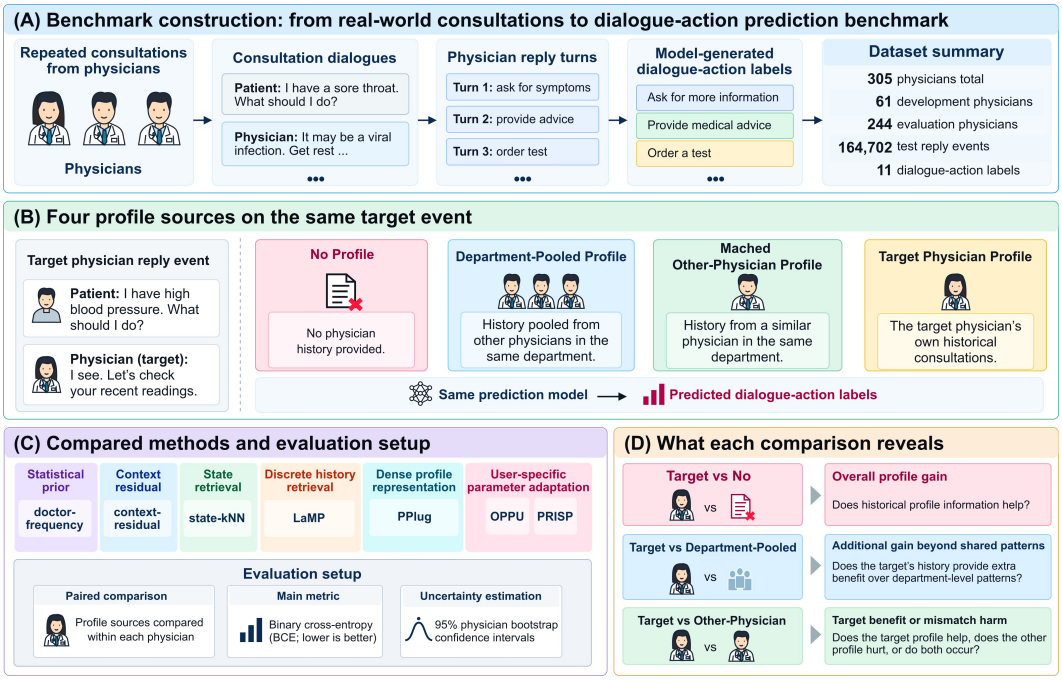


Figure 1: Benchmark design and evaluation. A: physician-reply prediction events and 11 functional dialogue-action labels are constructed from repeated consultations. B: the prediction event is held fixed while the source changes among no profile, a department-pooled profile, a matched other physician's profile, and the target physician's profile. C: seven methods output action probabilities under a common evaluation. D: the target profile is compared with no profile, pooled department history, and the matched other physician's profile.

3.2 Benchmark Construction

3.2.1 Data and Splits

We use the public Chunyu Yisheng Chinese online consultation dataset (Lu et al., 2022). We first retain consultations that can be linked to a physician and contain at least one physician reply, and remove exact duplicate dialogues. From the deduplicated data, we include physicians with at least 100 distinct consultations, providing enough consultations for profile construction and separate development and evaluation sets. Filtering yields 305 physicians, 115,906 consultations, and 1,337,748 physician-reply events across 14 departments. Each consultation contains multiple patient–physician turns. The experiments treat physician replies as prediction units and use the dialogue preceding the target reply as the current context. Appendix A.1 gives the full filtering procedure and reconciles the source-file record counts.

Of the 305 physicians, 61 are used for method development and the remaining 244 for evaluation. Within each physician, complete consultations are divided among PROFILE, DEVELOPMENT, HOLDOUT, and TEST in proportions of 60%, 15%, 10%, and 15%. All replies from one consultation stay in the same split, so profile histories and physician-reply events used for evaluation come from different consultations. The splits are consultation-disjoint rather than chronological: PROFILE is a separate reference set and is not guaranteed to precede the evaluation consultations. PROFILE supplies the profile histories. DEVELOPMENT consultations from the 61 development physicians are used for method selection, hyperparameter settings, and probability calibration. The main results are computed on TEST for the 244 evaluation physicians, while their DEVELOPMENT and HOLDOUT consultations are used to check stability across consultation subsets. TEST contains 164,702 physician-reply events from these 244 physicians. Appendix A.2 and Table A1 give the split sizes and integrity checks.

3.2.2 DIALOGUE-ACTION LABELS

We represent each target reply using 11 functional dialogue actions: information inquiry; requests for materials or test results; diagnostic assessment; medical explanation; treatment or medication guidance; testing or referral advice; lifestyle or self-care advice; safety-net or triage advice; emotional reassurance or support; follow-up or monitoring; and conversation management. A reply can contain multiple actions, so the task is multilabel prediction.

The source data do not include these labels. We construct supervision in two stages. A teacher model first labels a balanced sample of 5,975 events; these annotations then train a multilabel classifier that extends the labels to the remaining physician-reply events. Events labeled directly by the teacher retain those annotations, while the remaining events use classifier predictions. We select another 900 events for an inter-model consistency check. Appendix D describes the procedure and results.

3.3 FOUR-SOURCE PROFILE EVALUATION

3.3.1 FOUR PROFILE SOURCES

For each target physician, we construct reference histories from PROFILE and evaluate four sources on the same prediction event. Only the history source changes; the dialogue, prediction target, and other settings within each method remain fixed.

No Profile. No physician-level profile history is supplied. Only the method's shared parameters or global statistics remain.

Department-Pooled Profile. The source consists of PROFILE histories from other physicians in the target's department, excluding the target physician. Event-based methods construct a history set comparable in size to the target profile; statistical methods recompute their statistics using other physicians in the department.

Matched Other-Physician Profile. We select another physician in the same department whose PROFILE consultation count is closest to the target's and use that physician's complete PROFILE history. Unlike Department-Pooled, this condition supplies the coherent history of one specific other physician.

Target Physician Profile. The source is the target physician's own consultation history in PROFILE.

Appendix A.3 describes pooled-profile sampling and other-physician matching. Appendix A.4 and Table A2 report matching quality.

3.3.2 SOURCE COMPARISONS

We form five comparisons from the four sources. Target-versus-No measures the overall change after providing the target physician's profile. Department-Pooled-versus-No measures the change available from same-department history after excluding the target physician. Target-versus-Department-Pooled measures the additional improvement from the target physician's own history beyond shared department history.

Matched Other-Physician Profile supplies a different control: it preserves a complete, internally coherent individual history while replacing its owner with another physician from the same department and with a similar number of PROFILE consultations. Other-versus-No examines whether this alternative history helps or interferes with prediction, while Target-versus-Other measures the overall difference between the target and other physician's profiles. That difference can arise from improvement with the target profile or deterioration with the other profile, and is therefore interpreted alongside each profile's comparison with No Profile.

3.4 METHODS AND EVALUATION PROTOCOL

3.4.1 COMPARED METHODS

We select seven methods covering the main ways history enters a predictor: statistical aggregation, residual correction of the current context, state retrieval, discrete text retrieval, dense profile representations, and user-specific parameter adaptation (Table 1). Doctor-frequency, context-residual, and

state-kNN are simple baselines that test whether direct use of physician history already produces source differences. LaMP, PPlug, OPPU, and PRISP adapt retrieval, representation, and parameter-based approaches from personalized language modeling (Salemi et al., 2024; Liu et al., 2025; Tan et al., 2024; Park et al., 2026). Appendix B.1 and Table B1 describe their backbones, use of history, and task adaptations.

Table 1: Compared methods and their use of history.

Method	History use	Core mechanism
doctor-frequency	Statistical aggregation	Physician-level predictions from historical action frequencies, smoothed with department statistics
context-residual	Residual modeling	A current-dialogue prediction corrected using systematic historical residuals
state-kNN	State retrieval	Similar historical events retrieved from the specified source in a fixed state space
LaMP	Text retrieval	Relevant historical events supplied as language-model context
PPlug	Dense profile	Encoded history aggregated into a continuous profile representation
OPPU	Parameter adaptation	Source-specific adapters learned from the designated history
PRISP	Parameter adaptation	Source-specific parameters built from a task anchor and a small historical sample

3.4.2 Unified Source Adaptation

All seven methods are adapted to the same four-source evaluation, with source replacement limited to the component that uses history. Statistical methods re-estimate source-specific statistics, retrieval and representation methods switch historical records, and parameter-adaptation methods construct individual parameters from the designated profile. Within each method, the shared model, training settings, hyperparameters, and scoring procedure remain the same across the four conditions, and the output is an ordered vector of 11 action probabilities. Language-model methods use the same binary action-scoring and probability-calibration rules. Appendix B.2 describes the four-source implementation for each method, and Appendix B.3 describes language-model scoring and calibration.

3.4.3 Metrics and Statistical Analysis

Every method outputs 11 action probabilities. We therefore use binary cross-entropy (BCE) as the primary metric to evaluate the complete probability vector. Lower BCE indicates closer agreement between predicted probabilities and labels:

$$\mathrm{BCE} = -\frac{1}{11|\mathcal{I}|}\sum_{i\in\mathcal{I}}\sum_{k=1}^{11}\left[y_{ik}\log p_{ik} + (1-y_{ik})\log(1-p_{ik})\right].$$

We also report Brier score, Micro-F1, Macro-F1, and Primary Top-1 accuracy as supplementary metrics. Appendix B.4 defines these metrics, and Appendix Table C2 gives complete absolute results under all four sources. Absolute metrics are aggregated across events; the main source comparisons use physician-level paired analysis.

For physician d, let $L_d(c)$ be the mean BCE under source c. For any sources a and b, we first compute the paired difference within each physician:

$$\Delta_d(a,b) = L_d(b) - L_d(a),$$

where $\Delta_d(a,b)$ is the BCE improvement of source a over source b. We then average equally across the 244 evaluation physicians; positive values indicate lower BCE under source a.

We estimate 95% confidence intervals for the main source differences by physician-level bootstrap. Each resample draws 244 evaluation physicians with replacement, retains all target events from each sampled physician, and computes the mean paired difference. We repeat this 10,000 times and use

the 2.5th and 97.5th percentiles as interval endpoints. For Target-versus-Other, we also report the proportion of physicians with lower BCE under the target profile. Appendix Tables C1 and C3 give the complete differences and physician-level coverage.

4 RESULTS

Our main analysis compares the four sources within each method to distinguish changes associated with target-physician history, shared department information, and profile mismatch. Complete absolute metrics are in Appendix Table C2.

4.1 TARGET PROFILES IMPROVE PREDICTION FOR MOST METHODS

Six of the seven methods achieve lower BCE with the target physician's profile than with no profile (Figure 2A; Appendix Table C1). PPlug improves the most; doctor-frequency, context-residual, state-kNN, LaMP, and PRISP also show positive differences. OPPU is the exception: its average Target-versus-No change is close to zero, and the results do not establish a clear improvement from the target profile.

Overall, target-physician history usually improves prediction. Yet Target-versus-No captures the full change from providing that profile and does not show whether the gain depends on the history belonging to the target physician.

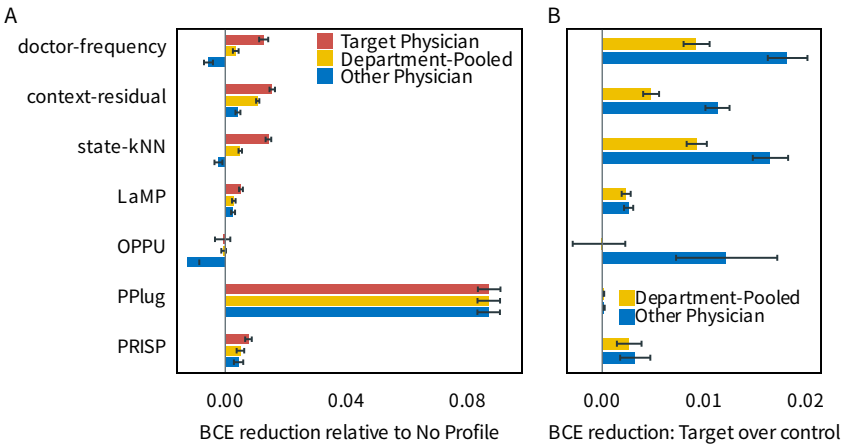


Figure 2: Five BCE comparisons among four profile sources on the test set. A: physician-mean differences for target, department-pooled, and other-physician profiles relative to no profile. The other physician is matched on department and PROFILE consultation count. B: physician-mean Target-versus-Department-Pooled and Target-versus-Other differences. Error bars are 95% physician-bootstrap confidence intervals. Positive values favor the first-named source.

4.2 SAME-DEPARTMENT HISTORY RETAINS PART OF THE PROFILE GAIN

For most methods, histories from other physicians in the same department retain a substantial part of the target-profile gain, even after excluding the target physician (Figure 3). The extent varies by method: PPlug retains almost the entire gain; context-residual, LaMP, and PRISP also retain a large proportion; doctor-frequency and state-kNN retain a smaller portion.

The target and pooled profiles also differ by method (Figure 2B). Switching to the target physician's history produces almost no further change for PPlug, some additional improvement for context-residual, LaMP, and PRISP, and larger gains for doctor-frequency and state-kNN.

For PPlug, almost all of the BCE reduction remains after the target physician is excluded. Doctor-frequency and state-kNN retain much less. Methods with similar Target-versus-No differences can therefore react quite differently when the history source changes.

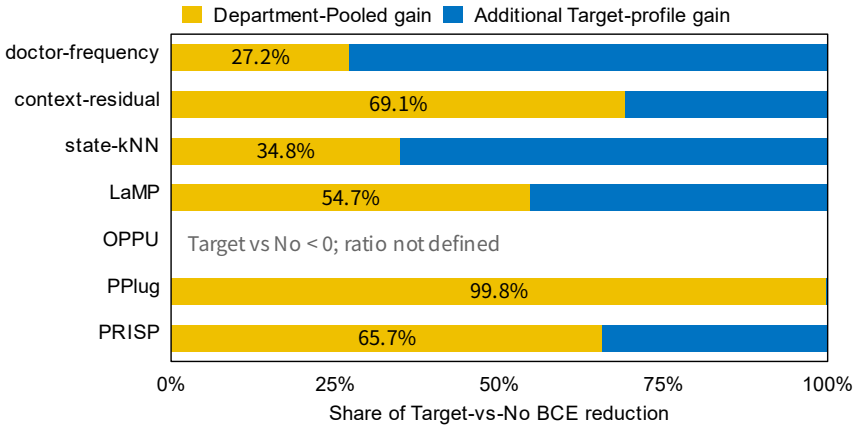


Figure 3: Proportion of the target-profile gain retained by the department-pooled profile. For methods with Target-versus-No > 0, the target profile's BCE reduction relative to no profile is set to 100%, and the plot shows how much of that reduction is already obtained with Department-Pooled. Yellow denotes the retained portion and blue the additional reduction obtained by switching to the target physician's profile. This ratio is calculated from the BCE point estimates for descriptive comparison and is not a causal decomposition of shared and physician-specific information.

4.3 Target Profiles Outperform Other-Physician Profiles for Different Reasons

Figure 4 places the physician-mean BCE changes for the target and other physician's profiles, each relative to no profile, in the same coordinate system. The Target-versus-Other differences correspond to different patterns across methods (Appendix Table C1).

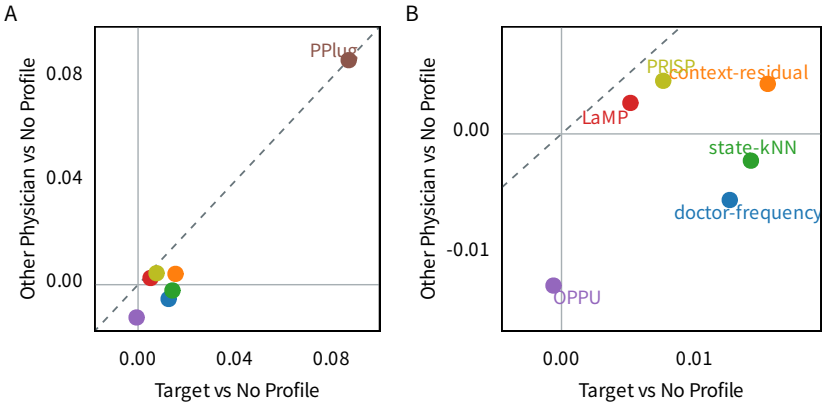


Figure 4: Physician-mean BCE differences for target and other-physician profiles, each relative to no profile. The dashed line marks equal changes. A shows the full scale; B enlarges the six methods other than PPlug. The axes show each profile's improvement over no profile.

OPPU shows no clear Target-versus-No improvement, while the other physician's profile yields higher BCE; its Target-versus-Other gap therefore comes mainly from the loss increase under the other profile. Doctor-frequency and state-kNN combine lower BCE under Target with higher BCE under Other. Context-residual, LaMP, and PRISP still achieve lower BCE under Other than under No Profile, but perform better under Target. PPlug changes almost identically under Target and Other, so its BCE gain is largely insensitive to which physician supplies the profile.

The Target-versus-Other BCE difference can thus reflect lower BCE under Target, higher BCE under Other, or both. No Profile and Other Physician provide different reference conditions.

4.4 Physician-Level Coverage and Stability across Consultation Splits

Target profiles outperform other-physician profiles on average, but physician-level coverage differs by method (Figure 5A; Appendix Table C3). Doctor-frequency, context-residual, and state-kNN improve for almost all evaluation physicians. Coverage is lower for LaMP and OPPU, while PPlug and PRISP have positive average differences without an advantage for every physician. The same positive mean difference can therefore be supported by nearly all physicians in one method and only a majority in another.

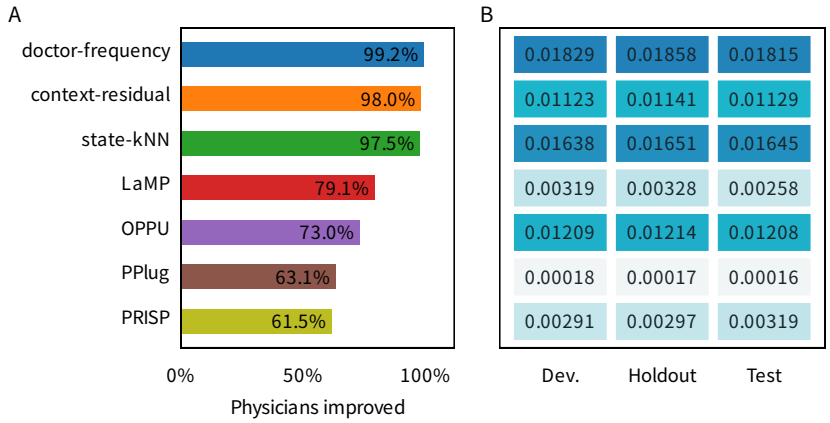


Figure 5: Physician-level Target-versus-Other results. A: proportion of the 244 test physicians with lower BCE under Target. B: physician-mean BCE differences on DEVELOPMENT, HOLDOUT, and TEST; darker shading marks larger differences.

Across the disjoint Development, Holdout, and Test consultations, all seven methods retain the same direction of the average Target-versus-Other difference, with small changes in magnitude (Figure 5B; Appendix Table C4). The result is stable across these consultation subsets.

5 Conclusions, Limitations, and Future Work

The seven methods use physician history in markedly different ways. PPlug improves almost equally with pooled, other-physician, and target profiles. OPPU shows no clear Target-versus-No gain and instead has higher loss under the other-physician profile. Doctor-frequency and state-kNN combine a target-profile gain with higher loss under another physician's profile. Evaluating all four sources on the same events distinguishes these cases, whereas Target-versus-No alone does not.

Limitations and Future Work. The benchmark is limited to one platform, one matched physician per target, and model-generated action labels. Future work can use multiple, more closely matched physicians, add an independently human-annotated subset, and test the evaluation across institutions, languages, and other personalization tasks.

Ethics Statement

This study uses the publicly released Chunyu Yisheng Chinese online consultation dataset to examine how profile sources affect physician dialogue-action prediction (Lu et al., 2022). The source repository states that the platform de-identified personal information and that the dataset publishers additionally removed sensitive messages manually. We did not attempt to recover patient or physician identities. The manuscript materials contain no raw consultation text, linkable physician profiles, or event-level labels. The results are intended for offline model evaluation, not physician performance assessment, automated accountability, or replacement of clinical judgment.

Reproducibility Statement

The anonymized supplementary material includes implementations of the profile-conditioned components used by the seven methods, evaluation code, figure-generation scripts, method configurations, and aggregate result tables. The files are organized by method, evaluation, plotting, configuration, and aggregate results, with usage instructions in the accompanying README. The main text and appendix describe profile construction, method adaptation, and statistical analysis.

AI Use Statement

Generative AI tools assisted literature search and organization, code development and debugging, and manuscript editing. The authors determined the research question, experimental design, method selection, and result interpretation, and verified all code, experimental results, citations, and main claims. AI-generated content served only as research and writing assistance, not as an independent source of evidence. AI systems are not authors.

References

David C. Chan, Matthew Gentzkow, and Chuan Yu. Selection with variation in diagnostic skill: Evidence from radiologists. *The Quarterly Journal of Economics*, 137(2):729–783, 2022. doi: 10.1093/qje/qjab048. URL https://doi.org/10.1093/qje/qjab048.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, volume 56 of *Proceedings of Machine Learning Research*, pp. 301–318, 2016. URL https://proceedings.mlr.press/v56/Choi16.html.

David Cutler, Jonathan S. Skinner, Ariel Dora Stern, and David Wennberg. Physician beliefs and patient preferences: A new look at regional variation in health care spending. *American Economic Journal: Economic Policy*, 11(1):192–221, 2019. doi: 10.1257/pol.20150421. URL https://doi.org/10.1257/pol.20150421.

Jannik Deuschel, Caleb Ellington, Yingtao Luo, Ben Lengerich, Pascal Friederich, and Eric P. Xing. Contextualized policy recovery: Modeling and interpreting medical decisions with adaptive imitation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10642–10660, 2024. URL https://proceedings.mlr.press/v235/deuschel24a.html.

Amy Finkelstein, Matthew Gentzkow, and Heidi Williams. Sources of geographic variation in health care: Evidence from patient migration. *The Quarterly Journal of Economics*, 131(4):1681–1726, 2016. doi: 10.1093/qje/qjw023. URL https://doi.org/10.1093/qje/qjw023.

Joseph Futoma, Michael Hughes, and Finale Doshi-Velez. POPCORN: Partially observed prediction constrained reinforcement learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3578–3588, 2020. URL https://proceedings.mlr.press/v108/futoma20a.html.

Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. DialMed: A dataset for dialogue-based medication recommendation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 721–733, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.60/.

Alihan Hüyük, Daniel Jarrett, and Mihaela van der Schaar. Inverse contextual bandits: Learning how behavior evolves over time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9506–9524, 2022. URL https://proceedings.mlr.press/v162/huyuk22a.html.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. doi: 10.1038/s41591-018-0213-5. URL https://doi.org/10.1038/s41591-018-0213-5.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In Katrin Erk and Noah A. Smith (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1094. URL https://aclanthology.org/P16-1094/.

Jiongnan Liu, Yutao Zhu, Shuting Wang, Xiaochi Wei, Erxue Min, Yu Lu, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. LLMs + persona-plug = personalized LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9373–9385, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.461. URL https://aclanthology.org/2025.acl-long.461/.

Xiaoxin Lu, Yubo Zhang, Jing Li, and Shi Zong. Doctor recommendation in online health forums via expertise learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1111–1123, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.79. URL https://aclanthology.org/2022.acl-long.79/.

David Molitor. The evolution of physician practice styles: Evidence from cardiologist migration. *American Economic Journal: Economic Policy*, 10(1):326–356, 2018. doi: 10.1257/pol.20160319. URL https://doi.org/10.1257/pol.20160319.

Chimaobi Okite, Naihao Deng, Kiran Bodipati, Huaidian Hou, Joyce Chai, and Rada Mihalcea. Benchmarking and improving LLM robustness for personalized generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16040–16072. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.870. URL https://aclanthology.org/2025.findings-emnlp.870/.

Junho Park, Dohoon Kim, and Taesup Moon. PRISP: Privacy-safe few-shot personalization via lightweight adaptation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24986–25003, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1146. URL https://aclanthology.org/2026.acl-long.1146/.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: A deep reinforcement learning approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 147–163, 2017. URL https://proceedings.mlr.press/v68/raghu17a.html.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7370–7392, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.399. URL https://aclanthology.org/2024.acl-long.399/.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, and Mausam. MediTOD: An English dialogue dataset for medical history taking with comprehensive annotations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16843–16877, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.936. URL https://aclanthology.org/2024.emnlp-main.936/.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2840–2861, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.167. URL https://aclanthology.org/2024.findings-acl.167/.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. Democratizing large language models via personalized parameter-efficient fine-tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6476–6491, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.372. URL https://aclanthology.org/2024.emnlp-main.372/.

Yuchen Wu, Edward Sun, Kaijie Zhu, Jianxun Lian, Jose Hernandez-Orallo, Aylin Caliskan, and Jindong Wang. Personalized safety in LLMs: A benchmark and a planning-based agent approach. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-3250. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8c2bb821410066459be64d03a4dc5719-Abstract-Conference.html.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. MedDialog: Large-scale medical dialogue datasets. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9241–9250, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.743. URL https://aclanthology.org/2020.emnlp-main.743/.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B. Cohen, and Emine Yilmaz. PersonaLens: A benchmark for personalization evaluation in conversational AI assistants. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18023–18055, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.927. URL https://aclanthology.org/2025.findings-acl.927/.

SUPPLEMENTARY MATERIALS

A BENCHMARK CONSTRUCTION DETAILS

A.1 DATASET FILTERING AND PHYSICIAN COHORT

We construct the benchmark from the public Chunyu Yisheng Chinese online consultation dataset (Lu et al., 2022). To ensure sufficient history for profile construction and separate consultations for development and evaluation, we first select physicians with consistent repeated consultation records. We retain records that can be linked to a physician, contain at least one physician reply, and belong to physicians with at least 100 distinct consultations, removing exact duplicate dialogues.

Filtering yields 305 physicians, 115,906 consultations, and 1,337,748 physician-reply events across 14 departments. Of these physicians, 61 are used for method development and 244 for evaluation.

During data reconciliation, the downloaded `dialogues.json` contained 119,133 consultation objects, whereas the official repository README[1] reported 119,128. The accompanying training, validation, and test TSV files contain 119,125 unique dialogue IDs, all present in the JSON file. Eight additional JSON dialogues have no corresponding TSV physician mapping.

A.2 SPLIT STATISTICS AND VALIDATION

Table A1 gives the full sizes of PROFILE, DEVELOPMENT, HOLDOUT, and TEST. We check data integrity after splitting. Each consultation belongs to only one split, and all physician replies in that consultation stay together. Exact duplicate dialogues do not cross splits. Each event uses only the continuous dialogue before the target reply; the target reply and later turns are excluded from the model input. Event IDs, physician IDs, consultation IDs, and split labels remain aligned. All eligible physicians have complete PROFILE, DEVELOPMENT, HOLDOUT, and TEST records.

Table A1: Split sizes. Consultation and all-event counts cover 305 physicians; evaluation events belong to the 244 evaluation physicians. Selection and calibration use only the 61 development physicians.

Split	Purpose	Consultations	All events	Evaluation events
PROFILE	Reference profiles	69,423	802,734	–
DEVELOPMENT	Method development (61 physicians); stability analysis (244 physicians)	17,238	198,898	161,337
HOLDOUT	Split stability	11,455	131,793	107,390
TEST	Main evaluation	17,790	204,323	164,702

A.3 PROFILE-SOURCE CONSTRUCTION DETAILS

For event-based methods, Department-Pooled histories are sampled without replacement from other physicians' PROFILE records in the same department, up to the number of events in the target physician's profile. Aggregate statistical methods instead recompute statistics directly from the same source set.

Matched Other-Physician assignments remain unchanged across DEVELOPMENT, HOLDOUT, and TEST. When multiple candidate physicians have the same absolute difference in PROFILE consultation count from the target, a fixed rule selects a unique match.

A.4 MATCHING QUALITY OF OTHER-PHYSICIAN PROFILES

All 244 evaluation physicians are matched within department. The absolute difference in PROFILE consultation count between target and matched physicians has a median of 3, an interquartile range

[1] https://github.com/polyusmart/Doctor-Recommendation

of 1–8, a 95th percentile of 23, and a maximum of 71. Exactly matching consultation counts occur in 13.9% of pairs.

Physician-turn event count is not a matching variable. Its absolute difference has a median of 617 and an interquartile range of 265–1,235. The fixed assignments use 202 distinct source physicians. Source reuse has a median of 1, a 95th percentile of 2, and a maximum of 3. There are no self-matches, missing mappings, invalid department-pooled sources, or pooled profiles below target capacity (Table A2).

Table A2: Matching quality of other-physician profiles (244 evaluation physicians).

Statistic	Value
Same-department pairs	100.0%
Exact match in PROFILE consultation count	13.9%
Absolute consultation-count difference, median [IQR]	3 [1, 8]
Absolute consultation-count difference, 95th percentile / maximum	23 / 71
Absolute physician-event-count difference, median [IQR]	617 [265, 1,235]
Distinct source physicians	202
Source reuse, median / 95th percentile / maximum	1 / 2 / 3
Self-matches / missing mappings	0 / 0
Invalid pooled sources	0
Pooled profiles below target event capacity	0.0%

B Methods and Evaluation Details

B.1 Method Implementations

The seven methods cover statistical aggregation, contextual residuals, state retrieval, text retrieval, dense profile representations, and parameter adaptation. Table B1 summarizes their backbones, history use, and adaptation.

Table B1: Backbones, history use, and adaptation for the seven methods.

Method	Category	Backbone	History use	Adaptation
doctor-frequency	Simple statistical	None	Action-frequency profile	None
context-residual	Simple statistical	Sparse text context model	Physician/source residual	Context model and residual
state-kNN	Simple retrieval	Fixed 20-dimensional state	State-neighbor retrieval	None
LaMP	Task-adapted published method	Qwen3-0.6B	Top-4 BM25 events	Task adaptation
OPPU	Task-adapted published method	Qwen3-0.6B	Source-specific adapter	Shared LoRA and source adapter
PPlug	Task-adapted published method	BGE-M3 + Qwen3-0.6B	Continuously weighted history and profile soft tokens	Alignment and shared modules
PRISP	Task-adapted published method	Mistral-7B-Instruct-v0.2	Source adapter and task anchor	Task anchor and source adapter

B.2 Four-Source Implementation by Method

The four conditions use the same target events and current dialogues and change only the historical source read by each method. Shared models, scoring procedures, and training settings unrelated to the source remain unchanged.

Doctor-frequency. Predictions are formed from historical action frequencies in the designated source. Target and Other use statistics from the target and matched physicians, respectively; Department-Pooled uses same-department statistics that exclude the target physician; and No Profile retains shared statistics without using one physician's history.

Context-residual. A base action probability is first estimated from the current dialogue and then adjusted using historical residuals from the designated source. Target, Other, and Department-Pooled estimate residuals from their respective histories, whereas No Profile uses the current-dialogue model without historical correction.

State-kNN. Under a fixed state representation, the method retrieves similar events from the designated PROFILE history and uses their residuals to adjust the current prediction. No Profile performs no physician-level historical retrieval.

LaMP. The method retrieves historical events most relevant to the current dialogue from the designated profile source and supplies them to the language model as additional context. No Profile uses the same prediction prompt without historical examples.

PPlug. A continuous profile representation is constructed from similarities between the current dialogue and the designated source history and then supplied to the language model. The three profile conditions differ only in history source; No Profile retains the shared task representation without an effective profile representation.

OPPU. A source-specific adapter is trained from the designated profile history on top of a shared task adapter. No Profile uses only the shared task adapter and loads no individual-source parameters.

PRISP. Source-specific parameters are constructed from the designated history on top of shared task parameters. No Profile retains the task-level parameters without individual adaptation.

Across the four conditions, each method changes only its historical statistics, retrieved records, profile representation, or source-specific parameters; all other settings remain the same.

B.3 Language-Model Scoring and Calibration

For LaMP, PPlug, OPPU, and PRISP, each of the 11 dialogue actions is formulated as a binary prediction task. Let $z_{ik}^{(m,c)}$ denote the difference between the scores for `1` and `0` produced by method m for event i, action k, and profile condition c. The calibrated probability is

$$\hat{p}_{ik}^{(m,c)} = \mathrm{clip}_{[10^{-5},\,1-10^{-5}]}\left[\sigma\Big(a_{mk} z_{ik}^{(m,c)} + b_{mk}\Big)\right], \qquad a_{mk} > 0. \tag{1}$$

Each language-model method fits its own calibrator, with a separate slope a_{mk} and intercept b_{mk} for each action. The parameters minimize binary cross-entropy on 37,561 Target Physician Profile events from DEVELOPMENT for the 61 development physicians. Once fitted, the parameters for a method and action remain unchanged across No Profile, Department-Pooled, Other, and Target.

B.4 Evaluation Metrics

All methods output probabilities for the 11 dialogue actions. In addition to BCE, we report Brier score, Micro-F1, Macro-F1, and Primary Top-1 to describe different aspects of prediction.

Brier score is the squared error between a predicted probability and its binary label, averaged over all events and action labels. Like BCE, it evaluates probabilistic predictions directly, but the two metrics penalize erroneous probabilities differently.

Micro-F1 pools true positives, false positives, and false negatives across all events and action labels before computing F1, making the result more influenced by frequent actions. **Macro-F1** computes F1

separately for each of the 11 actions and averages them, giving every action equal weight. For both metrics, each action probability is converted to a binary prediction at a fixed threshold of 0.5.

Primary Top-1 is the proportion of events for which the highest-probability predicted action matches the designated primary-action label.

Absolute BCE, Brier, Micro-F1, Macro-F1, and Primary Top-1 results are aggregated across events. The main source comparisons use physician-level BCE. Appendix Table C1 gives the complete source comparisons and confidence intervals; Appendix Table C3 gives physician-level coverage.

C Additional Results

C.1 Exact Profile-Source Comparisons

Table C1 gives exact values for the five main profile-source comparisons on TEST. Positive values indicate lower BCE for the first-named source. All differences are averaged with equal weight per physician. The Department-Pooled-versus-No and Other-versus-No intervals were computed after the main analysis from the frozen predictions, using the same physician-bootstrap procedure as the other three comparisons.

Table C1: Main profile-source comparisons on the test set. Positive values indicate lower BCE for the first-named source; physicians receive equal weight. Brackets give 95% physician-bootstrap confidence intervals.

Method	Target vs No	Department-Pooled vs No	Other vs No	Target vs Department-Pooled	Target vs Other
doctor-frequency	0.01263 [0.01117, 0.01417]	0.00343 [0.00251, 0.00440]	-0.00553 [-0.00703, -0.00408]	0.00920 [0.00798, 0.01054]	0.01815 [0.01625, 0.02014]
context-residual	0.01548 [0.01460, 0.01640]	0.01070 [0.01017, 0.01124]	0.00418 [0.00335, 0.00499]	0.00478 [0.00402, 0.00558]	0.01129 [0.01013, 0.01249]
state-kNN	0.01421 [0.01330, 0.01517]	0.00495 [0.00437, 0.00551]	-0.00224 [-0.00357, -0.00097]	0.00926 [0.00828, 0.01026]	0.01645 [0.01477, 0.01824]
LaMP	0.00517 [0.00454, 0.00581]	0.00283 [0.00222, 0.00341]	0.00259 [0.00191, 0.00322]	0.00234 [0.00191, 0.00279]	0.00258 [0.00214, 0.00304]
OPPU	-0.00060 [-0.00340, 0.00165]	-0.00053 [-0.00131, 0.00022]	-0.01269 [-0.01735, -0.00858]	-0.00007 [-0.00290, 0.00226]	0.01208 [0.00724, 0.01718]
PPlug	0.08705 [0.08336, 0.09082]	0.08691 [0.08324, 0.09068]	0.08688 [0.08321, 0.09064]	0.00013 [0.00008, 0.00019]	0.00016 [0.00010, 0.00024]
PRISP	0.00763 [0.00656, 0.00874]	0.00501 [0.00374, 0.00624]	0.00444 [0.00284, 0.00593]	0.00262 [0.00144, 0.00386]	0.00319 [0.00176, 0.00472]

C.2 Complete Metrics under Four Profile Sources

Table C2 gives the complete TEST metrics under the four profile sources. BCE, Brier, Micro-F1, Macro-F1, and Primary Top-1 are aggregated across events.

At the fixed 0.5 threshold, LaMP and PRISP have near-zero Micro-F1 and Macro-F1, while their BCE and Primary Top-1 remain nonzero. We retain these values as supplementary metrics; the main profile-source comparisons use physician-level paired BCE.

Table C2: Complete test metrics under the four profile sources. BCE, Brier, F1, and Primary Top-1 are aggregated across events.

Method	Profile condition	BCE	Brier	Micro-F1	Macro-F1	Primary Top-1
doctor-frequency	No Profile	0.3799	0.1169	0.1625	0.0910	0.2527
doctor-frequency	Department-Pooled	0.3763	0.1158	0.1804	0.1038	0.2855
doctor-frequency	Other Physician	0.3855	0.1185	0.1907	0.1121	0.2757
doctor-frequency	Target Physician	0.3665	0.1130	0.2129	0.1315	0.3114
context-residual	No Profile	0.3548	0.1089	0.2778	0.1646	0.3784
context-residual	Department-Pooled	0.3443	0.1054	0.3389	0.2123	0.3979
context-residual	Other Physician	0.3505	0.1072	0.3268	0.2053	0.3904
context-residual	Target Physician	0.3391	0.1040	0.3546	0.2304	0.4042
state-kNN	No Profile	0.3548	0.1089	0.2778	0.1646	0.3784
state-kNN	Department-Pooled	0.3494	0.1063	0.3319	0.2012	0.3979
state-kNN	Other Physician	0.3575	0.1077	0.3275	0.2015	0.3903
state-kNN	Target Physician	0.3398	0.1037	0.3532	0.2216	0.4066
LaMP	No Profile	0.4053	0.1253	<0.0001	0.0003	0.1767
LaMP	Department-Pooled	0.4021	0.1248	0.0001	<0.0001	0.1892
LaMP	Other Physician	0.4026	0.1249	0.0002	0.0001	0.1898
LaMP	Target Physician	0.3997	0.1240	0.0001	<0.0001	0.2021
OPPU	No Profile	0.3169	0.0955	0.4437	0.3069	0.4729
OPPU	Department-Pooled	0.3172	0.0957	0.4805	0.3461	0.4706
OPPU	Other Physician	0.3328	0.1012	0.4469	0.3232	0.4521
OPPU	Target Physician	0.3194	0.0968	0.4706	0.3456	0.4704
PPlug	No Profile	0.4717	0.1331	0.0093	0.0055	0.1904
PPlug	Department-Pooled	0.3825	0.1181	0.1381	0.0669	0.2930
PPlug	Other Physician	0.3825	0.1181	0.1383	0.0670	0.2928
PPlug	Target Physician	0.3824	0.1181	0.1385	0.0671	0.2931
PRISP	No Profile	0.4131	0.1280	0.0005	0.0003	0.1838
PRISP	Department-Pooled	0.4076	0.1262	<0.0001	<0.0001	0.1519
PRISP	Other Physician	0.4087	0.1265	<0.0001	<0.0001	0.1496
PRISP	Target Physician	0.4055	0.1257	<0.0001	<0.0001	0.1561

C.3 PHYSICIAN-LEVEL COVERAGE

Table C3 gives the proportion of TEST physicians with lower BCE under the target profile than under the other physician's profile.

Table C3: Physicians with lower BCE under Target than Other Profile

Method	Physicians with lower BCE under Target than Other Profile
doctor-frequency	99.2%
context-residual	98.0%
state-kNN	97.5%
LaMP	79.1%
OPPU	73.0%
PPlug	63.1%
PRISP	61.5%

C.4 STABILITY ACROSS CONSULTATION SPLITS

We calculate physician-mean Target-versus-Other BCE differences on the disjoint DEVELOPMENT, HOLDOUT, and TEST target-consultation sets. All three splits use the same 244 evaluation physicians and the same other-physician assignments, but different target consultations (Table C4).

Table C4: Physician-mean Target-versus-Other BCE differences across consultation splits.

Method	DEVELOPMENT	HOLDOUT	TEST
doctor-frequency	0.01829	0.01858	0.01815
context-residual	0.01123	0.01141	0.01129
state-kNN	0.01638	0.01651	0.01645
LaMP	0.00319	0.00328	0.00258
OPPU	0.01209	0.01214	0.01208
PPlug	0.00018	0.00017	0.00016
PRISP	0.00291	0.00297	0.00319

D SILVER DIALOGUE-ACTION LABELS

We represent physician replies using 11 functional dialogue actions: information inquiry; requests for materials or test results; diagnostic assessment; medical explanation; treatment or medication guidance; testing or referral advice; lifestyle or self-care advice; safety-net or triage advice; emotional reassurance or support; follow-up or monitoring; and conversation management. A reply can contain multiple actions.

The source consultations do not include these labels. We sample 5,975 events from PROFILE and DEVELOPMENT and annotate them with DeepSeek-v4-flash as the teacher model at a generation temperature of 0. To cover varied physician replies, sampling considers reply length, whether the reply contains a question, and action-related keywords. The teacher assigns functional actions using the current patient message and target physician reply and also selects a primary action. Its initial prompt distinguishes history/symptom inquiry from clarification requests; we merge these two categories into the final information-inquiry action before classifier training, yielding the 11-action ontology used in the benchmark.

The teacher annotations then train a multilabel classifier that extends the labels to the remaining physician-reply events. The current patient message and target reply are concatenated and represented with character-level TF–IDF features, followed by 11 one-vs-rest logistic-regression classifiers. Training and evaluation are grouped by physician, and the prediction threshold for each action is selected on a physician-grouped threshold-selection subset. On held-out physicians, the classifier reproduces the teacher annotations with a Micro-F1 of 0.6760 and a Macro-F1 of 0.6483. The

classifier labels the remaining 1,331,773 events, while the 5,975 teacher-annotated events retain the teacher labels. For classifier-generated events, the primary action is the highest-probability action among those predicted positive at their action-specific thresholds; if none crosses its threshold, the highest-probability action is used.

We additionally sample 900 events for separate annotation by DeepSeek, Qwen, and GLM without showing them the benchmark labels. Pairwise per-label agreement among the three models ranges from 0.949 to 0.958. These evaluations measure consistency with model-derived annotations; independent human validation of the action scheme remains incomplete.

The annotations were generated through the OpenCode Go Chat Completions endpoint. The teacher run used the service model alias `deepseek-v4-flash`; the audit used `deepseek-v4-flash`, `qwen3.7-plus`, and `glm-5`. All three audit models received the same instructions, label definitions, input fields, and output schema; only the model alias changed. These aliases are the identifiers retained in the run records, which do not include immutable provider-side revision identifiers. The complete teacher and audit prompts are included in the supplementary materials.

E Computational Resources and Software Environment

Experiments ran on a server with two Intel Xeon Gold 6330 CPUs and two NVIDIA A40 48 GB GPUs. Doctor-frequency, context-residual, and state-kNN ran on CPUs; each LaMP, OPPU, PPlug, and PRISP run used one A40 GPU, without multi-GPU distributed computation. The main software environment used Python 3.10.18, PyTorch 2.6.0+cu124, CUDA 12.4, and Transformers 4.57.3.