From Structured Data to Clinical Notes: Robust Clinical Decision Support with Fine-Tuned LLMs

Frederike Lübeck [1 2] **Jonas Wildberger** [2 3] **Frederik Träuble** [2 3] **Maximilian Mordig** [1 2]
Sergios Gatidis [4] **Andreas Krause** [1] **Bernhard Schölkopf** [1 2 3]

Abstract

Clinical machine learning models are typically trained on highly structured and consistent datasets but deployed in real-world settings dominated by unstructured clinical text, creating a fundamental challenge for practical adoption. In this work, we investigate whether large language models (LLMs), fine-tuned on structured patient data, can generalize effectively to unstructured clinical notes at inference time. Using the UK Biobank dataset for cardiovascular disease (CVD) risk prediction, we demonstrate that LLMs trained on structured representations achieve performance comparable to specialized tabular machine learning models. More importantly, we show that these models maintain strong predictive accuracy when applied to unstructured inputs, such as clinical notes, in both zero-shot and few-shot scenarios.

1. Introduction

Clinical decision-support systems powered by machine learning hold significant promise for improving clinical processes and patient outcomes, yet there is a stark mismatch between the nature of data used for model training and the data available during model deployment in real-world clinical practice. Large-scale biomedical databases such as the UK Biobank (Sudlow et al., 2015) and the All of US Research Program (All of Us Research Program Investigators et al., 2019) provide rich, structured datasets collected under controlled conditions. These datasets are ideally suited for the development of machine learning (ML) models due to their completeness, consistency, and standardized format. In contrast, the data available at the point of care in clinical settings presents a different picture: In the clinic, natural language text is the dominant form of documentation and information exchange. Clinical notes offer rich and nuanced patient information but pose challenges to traditional machine learning models as they depart from the regime of complete, structured, and standardized data assumed during training. This mismatch raises an important question: Can we leverage structured data during training and yet generalize to unstructured data inputs, such as clinical notes, during inference? To investigate this, we fine-tune large language models (LLMs) on structured clinical data from the UK Biobank, one of the largest biomedical databases containing detailed health information on over half a million individuals. We then evaluate these models in the challenging setting where the input format shifts to free-text clinical notes. We assess performance in a zero-shot scenario and explore adaptation using limited data from the target domain. We use cardiovascular disease (CVD) risk prediction as a test case—a task highly relevant for the clinic, as CVD remains a leading global cause of mortality (WHO, 2024) and early identification of high-risk individuals is an important step for effective prevention.

Our key contributions are: **(i)** We show that LLMs fine-tuned on structured data perform competitively with standard ML models in structured-data settings; **(ii)** We demonstrate that these models generalize well to unstructured inputs such as clinical notes in zero-shot and few-shot settings; **(iii)** We introduce a controlled testbed for studying structured-to-unstructured input shifts in clinical prediction tasks. Our findings suggest that LLMs, when fine-tuned on relevant clinical data, can provide a foundation for flexible, robust decision support—even when faced with the complex, unstructured data that dominates real-world clinical practice.

2. Method

We assess whether LLMs fine-tuned on structured clinical data can generalize to unstructured text for CVD risk prediction. Models are trained on structured inputs and tested on both structured and unstructured formats.

[1]Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland [2]Max Planck Institute for Intelligent Systems, Tübingen, Germany [3]ELLIS Institute, Tübingen, Germany [4]Stanford School of Medicine, CA, USA. Correspondence to: Frederike Lübeck <frederike.luebeck@inf.ethz.ch>.

Proceedings of the 1st ICML Workshop on Foundation Models for Structured Data, Vancouver, Canada. 2025.

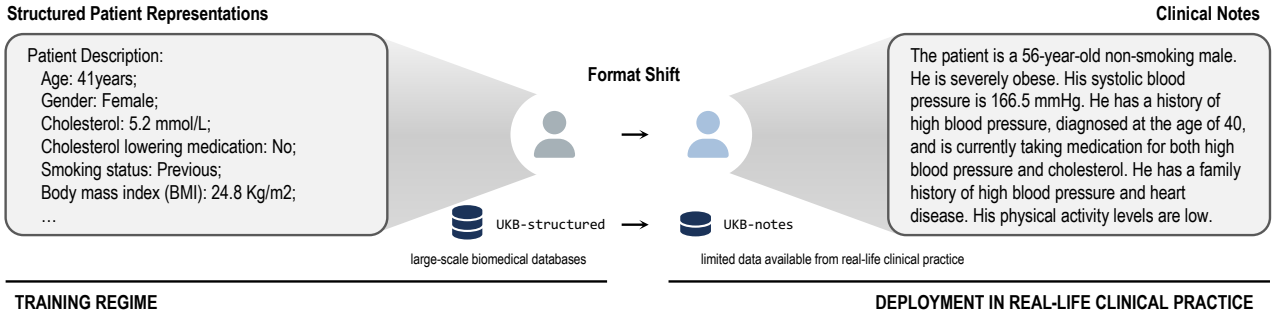

Figure 1. Illustration of the format shift from structured patient data to unstructured clinical notes, highlighting the discrepancy between training and deployment environments in clinical machine learning.

Task Definition We define our task to predict a patient's 10-year risk of developing CVD based on detailed health information available at a baseline assessment.

2.1. Data and Patient Representations

Our primary data source is the UK Biobank (Sudlow et al., 2015; UKB, 2025), a comprehensive biomedical database containing detailed health information from over 500,000 individuals. Beyond base cardiovascular risk factors, we defined nine information categories that capture broader aspects of patient health: Lifestyle & Environment, Sociodemographic factors, Physical Measures, Urine Assays, Blood Samples, Family History, Polygenic Risk Scores, ICD Codes, and Medical History.

We create two distinct types of patient representations on separate, non-overlapping splits of the dataset:

Structured Representations To create structured text representations (`UKB-structured`), we serialize patient data into detailed textual descriptions (see Fig. 1). Each patient profile is represented as a formatted, readable description that includes numeric values, categorical labels, and short textual items extracted from questionnaires. Feature names are expressed using medical terminology and descriptive labels to maximize informativeness. These structured descriptions serve as training inputs for our model.

Unstructured Clinical Notes In the absence of publicly available real-world datasets containing unstructured textual descriptions of patients with subsequent CVD outcomes, we leveraged LLMs to synthesize a corpus of free-text patient descriptions that mimic clinical notes (`UKB-notes`; see Fig. 1). Specifically, we prompt a separate LLM to produce clinical summaries based on each patient's structured features. This approach is informed by prior work on LLM-generated medical text (Agrawal et al., 2022; Van Veen et al., 2024). The resulting texts are diverse and significantly less structured. Overall, these summaries mostly preserved key clinical information, but they often expressed it in a more abstract or inferred form. For example, exact BMI values were sometimes replaced with phrases such as *the patient is obese*, and detailed physical activity metrics were summarized as *the patient is very active*. The focus of this work does not lie in evaluating these summaries. Instead, we treat them as given and assess how efficiently our model trained on structured representations can adapt to this unstructured input format. Examples and prompt templates are provided in Appendix §A.1.2 and §A.2.3.

2.2. Model Architecture and Training

Our approach follows the well-established pre-training and fine-tuning paradigm, utilizing a pre-trained transformer-based LLM with general language understanding capabilities as a foundation and tailoring it to the specific task of CVD risk prediction. We employ parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA; Hu et al. (2022)) on the `Mistral-7B-Instruct` model (Jiang et al., 2023). We fine-tune the model to predict a patient's CVD risk by framing the task as a binary classification problem in the token space (similar to Hegselmann et al. (2023); Belyaeva et al. (2023)). Instead of producing a numeric risk prediction in text form, we retrieve the likelihood of the model answering *Yes* or *No* to a question posed in binary form: *Will this patient experience a major cardiovascular event in the next ten years?* We extracted the logits and subsequently normalized them to generate the final CVD risk prediction. During training, we minimize the cross-entropy loss between predicted probabilities and observed outcomes. All fine-tuning is conducted on a cluster of NVIDIA A100 and H100 GPUs. This setup ensures that sensitive health data remains on-premise, fully preserving patient privacy. Moreover, because the UK Biobank dataset is not publicly available, we can be confident that there are no issues of data leakage and contamination, as this data was not used during the model's pre-training. Further details on the training process are provided in Appendix §A.1.

3. Experiments & Results

We evaluate our model under two complementary settings. First, we benchmark performance in an idealized setting using structured and complete patient data (`UKB-structured`), which mirrors the training distribution on a hold-out set and aligns with standard ML practice. Second, we evaluate the model on unstructured patient representations (`UKB-notes`), simulating the format shift encountered in real-world clinical deployment.

3.1. Idealized Setting: Structured & Complete Data

Baselines We compare our LLM-based model to two classes of baselines. The first group comprises established CVD risk models. These models were derived using Cox Proportional Hazards (CPH) models on large population cohorts across various geographic regions. These models use only a small set of risk factors as inputs, which we refer to as the *base risk factors*. The second group of baseline models comprises standard supervised ML models using tabular inputs. This group includes the CPH model (to parallel the methodology of the clinical scores), Logistic Regression (as a simple baseline), and Gradient Boosted Trees (instantiated by LGBM), which are widely used and known to perform strongly on tabular data. See Appendix §A.4.2 for details.

Using Only Base Risk Factors Figure 2 shows the performance using only the base risk factors as inputs. Both the LLM-based model and LGBM achieved state-of-the-art performance for CVD risk prediction, with an area under the receiver operating characteristic curve (AUROC) of 0.738. Performance was superior to simpler ML models, including logistic regression and the CPH model. Notably, all ML models surpassed established medical risk scores, which showed great variability in performance.

Incorporating Detailed Patient Information Figure 3 expands the input to include a broader range of patient information across the above-defined nine categories. Performance improves for both LGBM and LLM-based models as additional patient information is considered. The fine-tuned LLM performs on par with LGBM, which is notable given that LGBM is specifically designed and optimized for structured inputs, while the LLM operates on serialized text. Two feature categories reveal interesting differences. For blood sample data, which consists of continuous numerical measurements, LGBM slightly outperforms the LLM. In contrast, for ICD codes, which are sparsely occurring standardized codes for previous clinical conditions, the LLM performs better, likely due to its ability to leverage relationships between different diagnoses learned during pre-training.

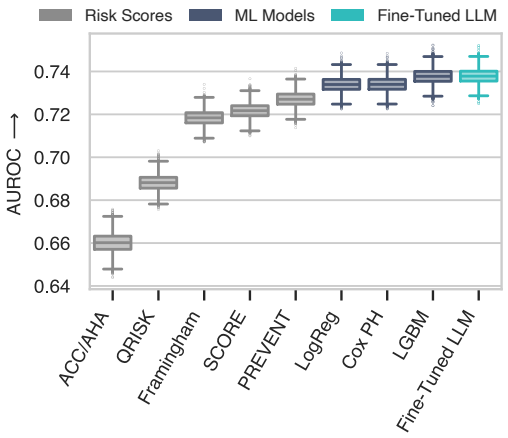

Figure 2. Comparison of CVD risk prediction models using only a limited set of base risk factors as inputs, across clinical risk scores, standard ML models, and the fine-tuned LLM using structured patient inputs. Predictive performance is measured by the area under the receiver operating characteristic curve (AUROC), reflecting the models' ability to distinguish between individuals who develop CVD and those who do not.

Simulating Incomplete Patient Records To assess robustness to missing data, we simulate incomplete patient records by selectively omitting all but one feature group at inference time while using the model trained on complete information for prediction. This allows us to examine how well the model performs under partial information, a common scenario in real-world clinical settings. As shown in Figure 3, the LLM-based model demonstrates greater resilience to missing inputs compared to LGBM. Note that since the LLM processes patients as textual descriptions, the absence of specific information simply results in a shorter prompt—there is no need for explicit imputation or placeholder values.

Overall, these results show that fine-tuned LLMs can accurately predict CVD risk using structured inputs, achieving similar performance to state-of-the-art tabular ML models, while demonstrating increased robustness to incomplete information.

3.2. Realistic Setting: Unstructured Clinical Notes

We evaluate the model under a realistic format shift: from structured patient descriptions to unstructured clinical notes. We compare the zero-shot performance and performance after adaptation via further fine-tuning. While structured datasets are typically large, real-world datasets with clinical notes are rare, making data efficiency a key consideration. We compare our approach to fine-tuning a pre-trained LLM from scratch only on the clinical notes.

As shown in Figure 4, the model fine-tuned on structured data performs well even in the zero-shot setting (AUROC 0.685) and improves further with minimal adaptation (0.697

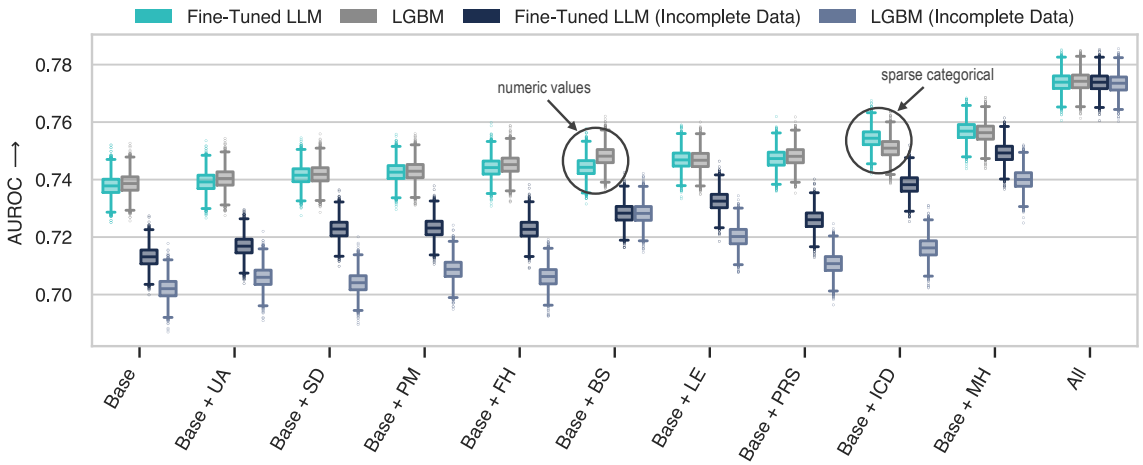

Figure 3. Predictive performance of the fine-tuned LLM and LGBM improves as additional patient information is incorporated. Acronyms: Urine Assays (UA), Sociodemographic factors (SD), Physical Measures (PM), Family History (FH), Blood Samples (BS), Lifestyle & Environment (LE), Polygenic Risk Scores (PRS), ICD Codes (ICD), and Medical History (MH). Each feature group includes the base risk factors. The *All* setting integrates all feature categories. Performance between the fine-tuned LLM and LGBM is competitive, with two exceptions (BS; ICD). Models denoted with *Incomplete Data* are trained using all features, but evaluated by omitting all but one feature group. The fine-tuned LLM shows more robust performance when confronted with missing values compared to LGBM.

with just 10 examples). In contrast, the model trained from scratch requires over 100 times more data (> 1000 points) to reach comparable performance. These findings demonstrate the strong generalization capabilities of the fine-tuned LLM and its ability to data-efficiently adapt to changed patient representations.

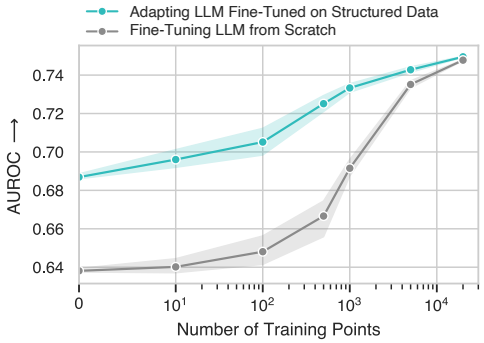

Figure 4. Evaluation of the fine-tuned LLM under a format shift to unstructured clinical notes, compared to an LLM fine-tuned only on clinical notes. The LLM fine-tuned on structured data achieves strong zero-shot performance and adapts efficiently with minimal additional data.

4. Discussion

Accurate disease risk estimation is central to preventive healthcare, yet the tools that clinicians rely on often fall short when confronted with the complexities of real-world clinical practice. These models are typically designed for clean, complete, and structured inputs—conditions that rarely hold outside of controlled research settings. Instead, natural language free-text is the dominant form of documentation in the clinic.

Our findings show that LLMs fine-tuned on structured patient data not only outperform established medical risk scores but also match the performance of specialized ML models optimized for tabular inputs. More importantly, we demonstrate that these models retain strong performance when confronted with real-world challenges, including missing data and unstructured input formats. We show that fine-tuning LLMs on structured representations enables robust generalization to free-text inputs, such as clinical notes, without requiring architectural changes or feature-specific engineering. Especially when considering comprehensive patient information—which we have shown to significantly improve performance—consistency in patient records is difficult to ensure. By describing patients in natural language, our approach moves beyond the rigidity of conventional methods that require fixed input features.

The results of this study suggest promising directions for further research, though several limitations should be noted. First, all training and evaluation was conducted on data from a UK-based research cohort that may not fully capture global demographic or clinical diversity. As such, this work should be viewed as a methodological proof of concept and is not intended for direct clinical use. Second, due to a lack of real-world datasets linking clinical notes to outcomes, we synthetically generated clinical notes. While prior work supports their validity, public datasets with authentic clinical text are needed to further advance this line of research.

In summary, our work presents a compelling pathway for using LLM-based models to bridge the gap between structured training data and the unstructured realities of everyday healthcare.

Acknowledgments

This work was carried out under UK Biobank application number 60520.

References

UK Biobank - UK Biobank. https://www.ukbiobank.ac.uk, January 2025.

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, August 2024.

Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., and Sontag, D. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H. F., and Van Der Schaar, M. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PLOS ONE*, 14(5):e0213653, May 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0213653.

All of Us Research Program Investigators, Denny, J. C., Rutter, J. L., Goldstein, D. B., Philippakis, A., Smoller, J. W., Jenkins, G., and Dishman, E. The "All of Us" Research Program. *The New England Journal of Medicine*, 381(7):668–676, August 2019. ISSN 1533-4406. doi: 10.1056/NEJMsr1809937.

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Muñoz, D., Smith, S. C., Virani, S. S., Williams, K. A., Yeboah, J., and Ziaeian, B. 2019 ACC/AHA Guideline on the Primary Prevention of Cardiovascular Disease: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation*, 140 (11), September 2019. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIR.0000000000000678.

Belyaeva, A., Cosentino, J., Hormozdiari, F., Eswaran, K., Shetty, S., Corrado, G., Carroll, A., McLean, C. Y., and Furlotte, N. A. Multimodal LLMs for health grounded in individual-specific data, July 2023.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. *Circulation*, 117(6):743–753, February 2008. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.107.699579.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., Yi, J., Zhao, W., Wang, X., Liu, Z., Zheng, H.-T., Chen, J., Liu, Y., Tang, J., Li, J., and Sun, M. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, March 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00626-4.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia,

J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The Llama 3 Herd of Models, November 2024.

Han, C., Kim, D. W., Kim, S., Chan You, S., Park, J. Y., Bae, S., and Yoon, D. Evaluation of GPT-4 for 10-year cardiovascular risk prediction: Insights from the UK Biobank and KoGES data. *iScience*, 27(2):109022, February 2024. ISSN 2589-0042. doi: 10.1016/j.isci.2024.109022.

Hegselmann, S., Buendia, A., Lang, H., Agrawal, M., Jiang, X., and Sontag, D. Tabllm: Few-shot classification of tabular data with large language models. In *International*

Conference on Artificial Intelligence and Statistics, pp. 5549–5581. PMLR, 2023.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., May, M., and Brindle, P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: Prospective open cohort study. *BMJ : British Medical Journal*, 335(7611):136, July 2007. ISSN 0959-8138. doi: 10.1136/bmj.39261.471806.55.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mistral 7B, October 2023.

Khan, S. S., Matsushita, K., Sang, Y., Ballew, S. H., Grams, M. E., Surapaneni, A., Blaha, M. J., Carson, A. P., Chang, A. R., Ciemins, E., Go, A. S., Gutierrez, O. M., Hwang, S.-J., Jassal, S. K., Kovesdy, C. P., Lloyd-Jones, D. M., Shlipak, M. G., Palaniappan, L. P., Sperling, L., Virani, S. S., Tuttle, K., Neeland, I. J., Chow, S. L., Rangaswami, J., Pencina, M. J., Ndumele, C. E., Coresh, J., and for the Chronic Kidney Disease Prognosis Consortium and the American Heart Association Cardiovascular-Kidney-Metabolic Science Advisory Group. Development and Validation of the American Heart Association's PREVENT Equations. *Circulation*, 149(6):430–449, February 2024. ISSN 0009-7322, 1524-4539. doi: 10.1161/CIRCULATIONAHA.123.067626.

SCORE2 working group and ESC Cardiovascular risk collaboration. SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, 42(25):2439–2454, July 2021. ISSN 0195-668X. doi: 10.1093/eurheartj/ehab309.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., and Collins, R. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLoS Medicine*, 12(3):e1001779, March 2015. ISSN 1549-1277. doi: 10.1371/journal.pmed.1001779.

Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., Grill, J.-B., Neyshabur, B., Bachem, O., Walton, A., Severyn, A., Parrish, A., Ahmad, A., Hutchison, A., Abdagic, A., Carl, A., Shen, A., Brock, A., Coenen, A., Laforge, A., Paterson, A., Bastian, B., Piot, B., Wu, B., Royal, B., Chen, C., Kumar, C., Perry, C., Welty, C., Choquette-Choo, C. A., Sinopalnikov, D., Weinberger, D., Vijaykumar, D., Rogozińska, D., Herbison, D., Bandy, E., Wang, E., Noland, E., Moreira, E., Senter, E., Eltyshev, E., Visin, F., Rasskin, G., Wei, G., Cameron, G., Martins, G., Hashemi, H., Klimczak-Plucińska, H., Batra, H., Dhand, H., Nardini, I., Mein, J., Zhou, J., Svensson, J., Stanway, J., Chan, J., Zhou, J. P., Carrasqueira, J., Iljazi, J., Becker, J., Fernandez, J., van Amersfoort, J., Gordon, J., Lipschultz, J., Newlan, J., Ji, J.-y., Mohamed, K., Badola, K., Black, K., Millican, K., McDonell, K., Nguyen, K., Sodhia, K., Greene, K., Sjoesund, L. L., Usui, L., Sifre, L., Heuermann, L., Lago, L., McNealus, L., Soares, L. B., Kilpatrick, L., Dixon, L., Martins, L., Reid, M., Singh, M., Iverson, M., Görner, M., Velloso, M., Wirth, M., Davidow, M., Miller, M., Rahtz, M., Watson, M., Risdal, M., Kazemi, M., Moynihan, M., Zhang, M., Kahng, M., Park, M., Rahman, M., Khatwani, M., Dao, N., Bardoliwalla, N., Devanathan, N., Dumai, N., Chauhan, N., Wahltinez, O., Botarda, P., Barnes, P., Barham, P., Michel, P., Jin, P., Georgiev, P., Culliton, P., Kuppala, P., Comanescu, R., Merhej, R., Jana, R., Rokni, R. A., Agarwal, R., Mullins, R., Saadat, S., Carthy, S. M., Cogan, S., Perrin, S., Arnold, S. M. R., Krause, S., Dai, S., Garg, S., Sheth, S., Ronstrom, S., Chan, S., Jordan, T., Yu, T., Eccles, T., Hennigan, T., Kocisky, T., Doshi, T., Jain, V., Yadav, V., Meshram, V., Dharmadhikari, V., Barkley, W., Wei, W., Ye, W., Han, W., Kwon, W., Xu, X., Shen, Z., Gong, Z., Wei, Z., Cotruta, V., Kirk, P., Rao, A., Giang, M., Peran, L., Warkentin, T., Collins, E., Barral, J., Ghahramani, Z., Hadsell, R., Sculley, D., Banks, J., Dragan, A., Petrov, S., Vinyals, O., Dean, J., Hassabis, D., Kavukcuoglu, K., Farabet, C., Buchatskaya, E., Borgeaud, S., Fiedel, N., Joulin, A., Kenealy, K., Dadashi, R., and Andreev, A. Gemma 2: Improving Open Language Models at a Practical Size, October 2024.

Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E. P., Seehofnerová, A., Rohatgi, N., Hosamani, P., Collins, W., Ahuja, N., Langlotz, C. P., Hom, J., Gatidis, S., Pauly, J., and Chaudhari, A. S. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30 (4):1134–1142, April 2024. ISSN 1546-170X. doi:

10.1038/s41591-024-02855-5.

WHO. Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), 2024.

A. Supplementary Material

Real-world clinical settings pose challenges to CVD risk prediction models, requiring them to handle diverse input information in varying formats and to adapt quickly to different healthcare environments. To address these requirements, we adopt the paradigm of task-specific fine-tuning of a pre-trained, general-purpose large language model (LLM). This paradigm has demonstrated key properties relevant to our setting: (i) transformer-based architectures enable flexible input representations via text prompts and depart from fixed, pre-defined input features; (ii) strong language understanding capabilities allow the incorporation of textual information; and (iii) efficient adaptation capabilities, e.g., through few-shot learning or fine-tuning.

We use a pre-trained LLM as a starting point and adjust it to the task of CVD risk prediction via supervised, parameter-efficient fine-tuning on real-world data.

A.1. Model Architecture and Fine-Tuning

Given a patient with individual-specific information, our model predicts the risk of developing CVD within the next 10 years. To achieve this, we fine-tuned LLMs for this specific task in a supervised manner using real-world data. This involves several key components: the choice of base LLM (see Section A.1.1), the construction of patient prompts (see Section A.1.2), the extraction of risk predictions (see Section A.1.3), and the supervised training process using parameter-efficient fine-tuning (see Section A.1.4).

A.1.1. Selecting Pre-Trained LLMs as Strong Starting Points

The LLMs we used are all autoregressive, decoder-only transformer models. We concentrated on open-access LLMs that we can deploy and fine-tune locally to ensure that no sensitive patient data leaves our servers. We focused on two classes of leading open-access LLMs for their balance between performance and computational efficiency during fine-tuning: small models (2-3 billion parameters) and medium-sized models (7-8 billion parameters). Specifically, we used `Mistral` (7B) (Jiang et al., 2023), `Llama` (3B, 8B) (Grattafiori et al., 2024), `Phi` (3B) (Abdin et al., 2024), and `Gemma` (2B) (Team et al., 2024). We use the instruction-tuned versions of these models. Since we observed similar performance after fine-tuning within each model class (see Fig. 5a), we continued with the `Mistral-7B-Instruct` model.

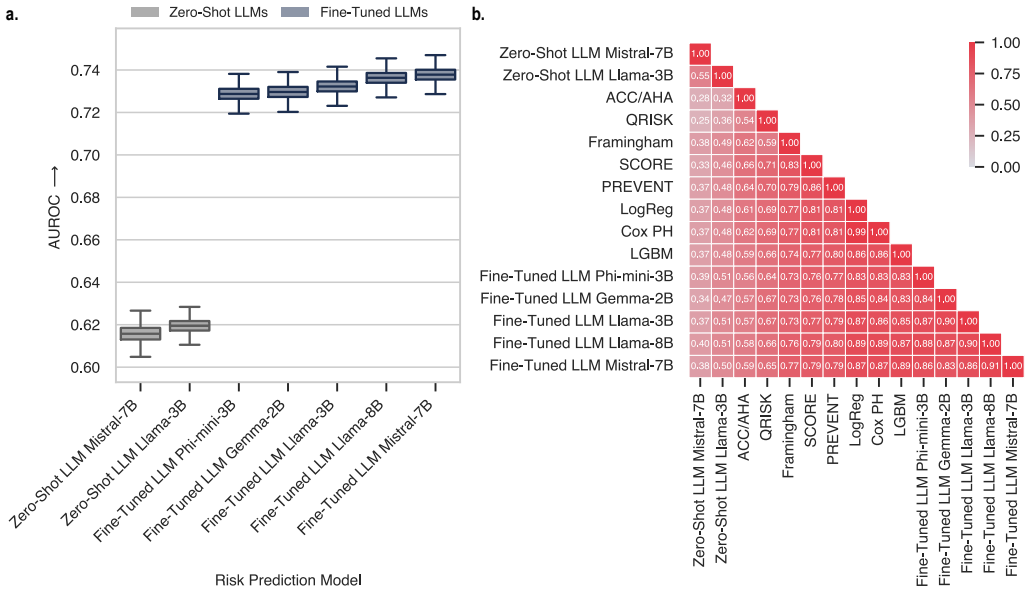

Figure 5. **Evaluation of risk prediction models using the base risk factors. a.** Comparison of different LLMs of small and medium size, both zero-shot and fine-tuned. LLMs not shown in the zero-shot group did not comply with the instructions. **b.** Correlation between predictions of different risk prediction models, as measured by the Kendall rank correlation coefficient.

A.1.2. Generating Patient Representations

We create two different types of natural language prompts for describing patients: a highly structured one (for `UKB-structured`) and an unstructured text that simulates clinical notes (for `UKB-notes`). Here, we describe how we generated the prompts that we used as input to our model. The data sources used for this process are described in Section A.3.

Structured Representations To create structured text representations, we serialize data on patients into a detailed textual description, as shown below.

Structured patient representations

```
Patient description:
   Gender: Male;
   Age: 41 years;
   <Feature Name>: <Feature Value>;
   ...
```

For all features, we use descriptive and precise names. Depending on the type of feature, the value can be a number (rounded to 1 digit), a short text snippet derived from questionnaire-type information, or a list thereof for questions that allow multiple answers.

Textual Representations In the absence of real-world datasets containing unstructured textual descriptions of patients with corresponding 10-year CVD outcomes, we leveraged LLMs to generate patient descriptions that mimic realistic clinical notes. For this, we followed prior work demonstrating LLMs' effectiveness in generating realistic medical summaries (Agrawal et al., 2022; Van Veen et al., 2024). We used structured patient information as the input and instructed the model to produce a free-text summary of each patient. For the generation, we used two different system prompts, shown below:

Prompt I for generating patient summaries

```
You are a medical doctor writing detailed clinical notes.

Patient description:
   <Feature Name>: <Feature Value>;
   ...

Based on this information, generate a concise and natural clinical summary
describing the patient in a few sentences.
```

Prompt II for generating patient summaries

```
You are a medical doctor writing detailed clinical notes.

Patient description:
   <Feature Name>: <Feature Value>;
   ...

Based on this information, generate a brief summary of the patient with an
emphasis on relevant cardiovascular-related information. Do not provide
risk evaluation or any clinical judgment.
```

A.1.3. Binary Classification in the Token Space

To fine-tune LLMs for CVD risk prediction, we framed the problem as a binary classification task in the token space (similar to (Hegselmann et al., 2023; Belyaeva et al., 2023)). Instead of producing a numeric risk prediction in text form, we retrieved

the likelihood of the model answering *Yes* or *No* to a question posed in binary form: *Will this patient experience a major cardiovascular event in the next ten years?* We extracted the logits and subsequently normalized them to generate the final CVD risk prediction. During training, we completed the prompt with a binary label based on the true observed 10-year CVD outcome of each patient and learned the parameters to minimize the cross-entropy loss between predicted probabilities and observed outcomes.

A.1.4. Efficient Fine-Tuning via Low-Rank Adaptation (LoRA)

Given the high computational cost of fully training such large models, we employed parameter-efficient fine-tuning (PEFT; (Ding et al., 2023)), namely Low-Rank Adaptation (LoRA; (Hu et al., 2022)). LoRA introduces lightweight adapter modules to the attention blocks of the transformer model while keeping the original pre-trained parameters frozen. Specifically, we targeted the query, key, and value projection layers, with a rank value of 16. With this approach, we updated only around 0.13% of the model parameters during fine-tuning and thereby significantly reduced computational demands. The training was done on a cluster of NVIDIA H100 and A100 GPUs.

A.2. Base Model and Model Adaptation

A.2.1. Fine-Tuning on Structured Data

Following the fine-tuning process outlined in Section A.1.4, we developed our base model using structured representations of patients from the UK Biobank (`UKB-structured`). We train the model for two epochs on all patients from the training dataset ($n = 467$k), using mini-batches of size 8-16[1]. The hyperparameters were chosen based on the model's performance on the validation set.

To assess the importance of different patient information for risk assessment, we trained expert models for each of the 10 information groups defined below (see Section A.3 for details), each focusing on a different aspect of health-related patient information. For this, we generated patient descriptions solely using the information contained in the specific feature group and the base risk factors. Hence, this process resulted in 11 different expert models: Base, using only the base risk factors; Base+X for the 9 different feature groups; and AllPatientInfo, which uses information from all feature groups simultaneously. Each model was specifically designed to deal with a fixed feature group at inference time.

A.2.2. Handling Incomplete and Variable Patient Information

To evaluate the model under incomplete data, we used the AllPatientInfo, i.e., the model trained on complete information of all feature groups, and provided incomplete information during inference. Note, however, that missing values are not explicit `null` values that require imputation, e.g., with the population median. Instead, incomplete information is only implicit and is simply left out of the patient descriptions.

A.2.3. Adapting to Textual Patient Representations

Our initial model was fine-tuned at scale on structured patient representations. Even though these representations were encoded in text format, they followed a highly standardized and consistent structure. In contrast, real-world clinical settings rarely provide such uniformity. Patient information is often documented in unstructured formats, such as clinical notes, physician reports, or discharge summaries, making free-text one of the most prevalent data modalities in practice. A key challenge for CVD risk prediction models is thus the ability to process and reason over unstructured text inputs. Therefore, we conducted an experiment in which we evaluate how well the model trained exclusively on structured inputs generalizes to unstructured text representations in a zero-shot setting. Additionally, we examine how efficiently it can be adapted to this new input format via further fine-tuning. For comparison, we also fine-tune the base LLM directly on the textual patient descriptions *from scratch*, without any prior fine-tuning on structured data.

We perform this experiment on our generated dataset of clinical notes (`UKB-notes`). For some prompts, we provided only the base risk factors as inputs (using Prompt I), and for others, we provided more detailed patient information (all feature groups except lab values, i.e., UA and BS; using Prompt II). We generate this dataset for a subset ($n = 40\,000$) of the UK Biobank cohort. Importantly, we use data from patients not seen during the first fine-tuning stage. To assess the data efficiency, we randomly select subsets for training using different random seeds.

[1]The batch size varied depending on the length of the patient descriptions across different settings.

The generated patient summaries averaged 135 tokens with base risk factors and 248 tokens with additional patient information. We capped the lengths at 200 tokens (cropping 3% of cases) for base summaries and 400 tokens (cropping 11% of cases) for detailed ones.

Manual inspection of a subset of the generated summaries confirmed that relevant clinical information was generally preserved, though often rephrased. For example, numerical values were replaced with qualitative descriptors (e.g., *elevated cholesterol levels*), and some features were inferred indirectly (e.g., mentioning obesity instead of stating the BMI value). Summaries based on base risk factors retained nearly all original information, while those including more granular inputs (e.g., physical activity broken down by type and duration) tended to be abstracted (e.g., *the patient is very active*). Our focus in this work does not lie in evaluating these summaries. Instead, we treat them as given and examine how effectively LLMs can learn from such text-based inputs and how efficiently a model trained on structured data adapts to this unstructured format. Examples of such patient summaries can be found below.

A.2.4. Examples of textual patient representations

> The patient is a 41-year-old non-smoking, non-diabetic female of white ethnicity with a BMI of 23.1 Kg/m2. She has a cholesterol level of 4.9 mmol/L and an HDL cholesterol level of 1.9 mmol/L. Her blood pressure, as measured automatically, is 108.5 mmHg. She is not currently taking any cholesterol-lowering medication or blood pressure medication. Her eGFR is 120.37, indicating normal kidney function.

> The patient is a 61-year-old male with a BMI of 24.9 Kg/m2, previously a smoker but not currently. He has a cholesterol level of 4.8 mmol/L and a low HDL cholesterol level of 1.1 mmol/L. His systolic blood pressure is 133 mmHg. He does not have diabetes, is not on blood pressure medication, and does not take cholesterol-lowering medication. His estimated glomerular filtration rate (eGFR) is 77.55, indicating good kidney function.

> The patient is a 48-year-old non-smoking, non-diabetic female of white ethnicity with a normal body mass index (BMI) of 21.9 Kg/m2. She has a borderline high cholesterol level, with a low HDL cholesterol level. Her blood pressure, as measured automatically, is slightly elevated at 134.5 mmHg. She is not currently on any cholesterol-lowering medication or blood pressure medication. Her estimated glomerular filtration rate (eGFR) is within the normal range at 96.05.

> The patient is a 41-year-old female with a BMI of 23.1 Kg/m2, who has never smoked and has no history of diabetes or hypertension. Her cholesterol level is 4.9 mmol/L, with an HDL cholesterol of 1.9 mmol/L. She is currently not on any cholesterol-lowering medication. Her systolic blood pressure, as measured automatically, is 108.5 mmHg. She has a family history of non-accidental death in close genetic family members. Her PRS for cardiovascular disease (CVD) is 2.3 relative risk, and her PRS for venous thromboembolic disease (VTE) is 2.2 relative risk. She engages in regular walking and light DIY, and her sleep duration is 8 hours/day. She consumes alcohol three or four times a week, with an average weekly spirits intake of 4 measures. Her maximum workload during a fitness test was 80 Watts, and her maximum heart rate during the test was 139 bpm. She lives in a house or bungalow with 2 people and has a college or university degree as her highest qualification.

> The patient is a 65-year-old female with a BMI of 22.9 Kg/m2. She is a current smoker and has a systolic blood pressure of 122 mmHg. Her cholesterol level is 6.1 mmol/L, with an HDL cholesterol of 1.3 mmol/L. She is not taking any cholesterol-lowering medication. Her estimated glomerular filtration rate (eGFR) is 89.92. She has a standard polygenic risk score (PRS) for coronary artery disease (CAD) of 1.1 relative risk. She has no history of diabetes, hypertension, or cardiovascular disease. She is physically active, walking 7 days a week and engaging in moderate physical activity for 300 minutes a day. She has no known vascular or heart problems diagnosed by a doctor. Her sleep duration is 6 hours a day, and she does not snore or daytime doze. She has a standard PRS for hypertension of 0.3 relative risk.

A.3. Data Sources and Cohort Descriptions

The UK Biobank (UKB, 2025; Sudlow et al., 2015) ($n = 467\,063$) serves as the main dataset for training, adaptation, and evaluation.

The UK Biobank is a large-scale longitudinal biomedical database containing detailed health information of over approximately half a million individuals across the UK. It offers a comprehensive repository of patient characteristics, encompassing sociodemographic information, physical measures, lab values, genetic data, lifestyle factors, medical history, and more. Information was collected at a baseline assessment, and after that, disease outcomes and mortality were continuously recorded in a follow-up period of up to 19 years.

Task & Outcome Definition We define our task as predicting the risk of developing a fatal or non-fatal CVD event within 10 years of the baseline assessment. Hereby, a CVD event is defined as the first occurrence of any of the following ICD-9 and ICD-10 diagnosis codes:

- **ICD-9:** 410–414 (ischemic heart diseases), 430–434 (hemorrhagic and ischemic stroke), and 436–438 (cerebrovascular diseases)
- **ICD-10:** F01 (vascular dementia), I20–I25 (ischemic heart diseases), I50 (heart failure), and I60–I69 (cerebrovascular diseases)

This aligns with definitions used in prior studies (Alaa et al., 2019; D'Agostino et al., 2008). We combined information from three sources: hospital in-patient admissions, self-reported data, and death registries. Participants with a history of CVD prior to the baseline assessment ($n = 35\,070$) were excluded, applying the same definition for CVD as used for the outcome variable.

Cohort The final cohort comprised $467\,063$ participants aged 37–73 years at baseline. The cohort was randomly split into a training (75%), test (20%), and validation set (5%). All reported results are computed on the test set unless stated otherwise. Over the 10-year follow-up period, 7.5% ($n = 34\,983$) of the participants developed CVD. Table 1 shows the baseline characteristics of the study population.

Comprehensive Health Information We incorporate comprehensive health-related information on individuals and have defined ten distinct information categories designed to reflect realistic clinical scenarios.

- **Base Risk Factors (Base):** This set of features is commonly used by established CVD risk scores. It consists of age, gender, smoking status, diabetes, total cholesterol, HDL cholesterol, cholesterol medication use, blood pressure, blood pressure medication use, body mass index (BMI), ethnic background, and estimated glomerular filtration rate (eGFR).
- **Polygenic Risk Scores (PRS):** These values quantify the genetic susceptibility of an individual to a broad range of diseases and traits by aggregating the effects of multiple genetic variants. It covers conditions such as cardiovascular diseases, different cancer types, autoimmune disorders, metabolic traits, and neurological and psychiatric disorders. We include 36 scores.
- **Medical History (MH):** Self-reported health information collected through questionnaires, encompassing diagnosed conditions with the individual's age at diagnosis, past medical procedures, medication use, and screening history.
- **Blood Samples (BS):** 43 laboratory-analyzed biomarkers measured in the blood sample collected at recruitment, including 26 biochemistry markers and 17 haematological assays.
- **Family History (FH):** Questionnaire-based information on health conditions of biological and adopted family members, offering insights into hereditary health risks.
- **Lifestyle and Environment (LE):** Self-reported data on physical activity, sleep habits, smoking behavior, and alcohol consumption, providing a comprehensive view of daily routines, health behaviors, and environmental exposures.
- **Physical Measures (PM):** Measurements of body size, body composition by impedance, electrocardiogram (ECG) during exercise, arterial stiffness, and spirometry.

- **Sociodemographics (SD):** Information on living arrangements, household composition, income, education level, employment status, and work conditions.
- **Urine Assays:** Biochemical measurements of urinary components, including creatinine, microalbumin, potassium, and sodium.
- **ICD Codes (ICD):** A record of all past diagnoses using ICD-9 and ICD-10 codes.

We provide the exact list of field IDs and features used per category in Table 2.

Table 1. Characteristics of the UK Biobank Cohort, excluding participants with CVD prior to the baseline assessment. We report median values and their standard deviation.

	Female (n = 261 030)	**Male (n = 206 033)**
Age (years)	57.00 (7.99)	57.00 (8.21)
BMI (kg/m²)	26.03 (5.14)	27.18 (4.17)
Total Cholesterol	225.99 (43.09)	214.66 (42.13)
HDL Cholesterol	60.33 (14.58)	48.26 (12.03)
Systolic Blood Pressure	133.00 (19.23)	139.50 (17.38)
Blood Pressure Medication	15.80%	19.76%
eGFR	97.60 (13.01)	97.53 (12.69)
Smoker	8.78%	12.39%
Diabetic	3.37%	5.74%

Table 2. List of field IDs used for the information categories in the UK Biobank. IDs marked with an asterisk are further processed into features. Information on the field can be found on the UK Biobank Showcase Webpage.

	Field IDs
Base	31, 93, 2443, 4080, 6153*, 6177*, 20116, 21000, 21001, 21003, 30690, 30700*, 30760
PRS	26202, 26204, 26206, 26210, 26212, 26214, 26216, 26218, 26220, 26223, 26225, 26227, 26229, 26232, 26234, 26238, 26240, 26242, 26244, 26246, 26248, 26250, 26252, 26254, 26258, 26260, 26265, 26267, 26269, 26273, 26275, 26278, 26283, 26285, 26287, 26289
MH	2178, 2188, 2296, 2306, 2316, 2345, 2355, 2365, 2415, 2443, 2453, 2463, 2473, 2492, 2844, 2966, 2976, 3005, 3761, 3786, 3809, 3992, 4012, 4022, 4041, 4717, 6150, 6151, 6152, 6153, 6154, 6155, 6177, 6179
BS	23000, 30000, 30010, 30020, 30030, 30040, 30050, 30060, 30070, 30080, 30090, 30100, 30110, 30120, 30130, 30140, 30150, 30160, 30600, 30610, 30620, 30630, 30640, 30650, 30660, 30670, 30680, 30690, 30700, 30710, 30720, 30730, 30740, 30750, 30760, 30770, 30780, 30790, 30810, 30840, 30860, 30870, 30880, 30890
ICD	41280*, 41270*, 41281*, 41271*
FH	1807, 1845, 3526, 4501, 20107, 20110, 20111, 20112, 20113, 20114
SD	670, 709, 728, 738, 767, 777, 796, 806, 816, 826, 845, 3426, 4674, 6138*, 6143, 20119
LE	864, 874, 884, 894, 904, 914, 924, 943, 971, 981, 991, 1001, 1011, 1021, 1070, 1080, 1090, 1160, 1190, 1200, 1210, 1220, 1239, 1249, 1259, 1269, 1279, 1558, 1568, 1578, 1588, 1598, 1608, 1618, 1628, 2624, 2634, 3637, 3647, 20116, 20117, 20160, 20161, 20162, 22035, 22036, 22037, 22038, 22039
PM	3062, 3063, 3064, 4194, 4195, 4196, 4198, 4199, 4200, 4204, 4207, 5983, 6015, 6016, 6017, 6032, 6033, 6034, 6039, 20150, 20151, 20256, 20257, 20258, 21001, 21021, 23098, 23099, 23100, 23101, 23102
UA	30500, 30505, 30510, 30520, 30525, 30530, 30535

A.4. Evaluation

A.4.1. Metrics

We evaluated the models using standard metrics suitable for unbalanced binary classification tasks. Specifically, we used the area under the receiver operator curve (AUROC) to assess the model's ability to differentiate between individuals who develop the disease and those who do not. For all metrics, we report the median value and their 95% spread across 5000 bootstrapping rounds. The observed large spreads are a result of high sample dependence, which is likely due to class imbalance. We decided not to measure randomness across different training runs (e.g., different seeds or initialization parameters) due to the high computational cost. However, we observed very stable results with respect to such randomness.

A.4.2. Comparisons & Baselines

For all experiments using tabular input features, we compared our method with various baseline methods, including medical risk scores, standard machine learning methods, and LLMs (zero-shot).

Medical Risk Scores We implemented medical risk scores derived from different geographic cohorts. We list all risk scores and their geographic regions in Table 3.

Table 3. Medical Risk Scores

Risk Score	Derivation Cohort
Framingham (D'Agostino et al., 2008)	US
PREVENT (Khan et al., 2024)	US
ASCVD (AHA/ACC) (Arnett et al., 2019)	US
SCORE2 (SCORE2 working group and ESC Cardiovascular risk collaboration, 2021)	Europe
QRISK (Hippisley-Cox et al., 2007)	UK

Machine Learning Baselines The second group of baseline models comprises standard supervised machine learning methods, including the Cox Proportional Hazards model, logistic regression, and gradient-boosted trees. We used the following software packages for the implementations: `lifelines` for the Cox PH model, `sklearn` for logistic regression, and `lightgbm` for gradient-boosted trees.

LLMs (Zero-Shot) To assess the zero-shot predictions of different pre-trained LLMs, we provided a patient description, gave precise instructions, and extracted the prediction from the response, similar to (Han et al., 2024). We instructed the models to utilize a JSON format within their responses to ensure straightforward extraction of the numeric risk prediction. Specifically, our instruction was: *Based on the provided patient description, what is the estimated 10-year risk of cardiovascular disease (CVD)? Please provide your answer solely as a numeric percentage in a machine-readable JSON format.* We generated 100 new tokens and extracted the risk prediction from the response. If no valid JSON was provided, we set the prediction to `nan`. When a model did not comply with the instructions, all predictions were invalid and hence, we were not able to compute any evaluation metrics.