Contrastive Language-Genotype Pre-training Facilitates Generalizable Disease Risk Prediction

Anonymous ACL submission

Abstract

We propose Contrastive Phenotype–Genotype Pretraining (CPGP), a CLIP-style framework that uses longitudinal ICD-10 trajectories as semantic supervision to learn transferable polygenic risk scores-derived genetic representations by aligning each participant's genetic profile with a language- and time-aware clinical history embedding. We evaluate CPGP through zero-shot cross-modal retrieval, cross-disease prediction, and agentic 10-year disease risk estimation. These settings test genotype–phenotype alignment, transferable genetic prediction beyond PRS baselines, and the utility of CPGP retrieval for future label-limited risk estimation. Experiments on UK Biobank show that CPGP improves zero-shot genetic-to-clinical retrieval with scale, outperforms PRS baselines across diverse disease endpoints, and provides useful evidence for modern agentic reasoning. These results suggest that clinical language trajectories can serve as scalable supervision for phenotype-aware genetic representation learning.

1 Introduction

Aligning clinical language with human genetics is an emerging frontier for biomedical natural language processing. However, connecting the dots between thousands of traits and millions of genetic variants is a daunting task. During the past two decades, significant progress has been made using genome-wide association study (GWAS) (Stranger et al., 2011; Welter et al., 2014) and polygenic risk score (PRS) (Torkamani et al., 2018; Lewis and Vassos, 2020). Despite great success, most existing PRS models are developed in a disease-specific manner, which requires results from well-powered GWAS conducted *a priori* (Uffelmann et al., 2021; Abdellaoui et al., 2023). Not only costly and time-consuming, developing thousands of isolated PRS models independently fails to capture complex correlations among phenotypes and ignore comorbidities and other underlying health conditions of patients, thus limiting their predictive performance (Kullo, 2025; Wang et al., 2022).

Rather than modeling each disease independently, large-scale biobanks such as the UK Biobank (Bycroft et al., 2018) make it possible to study genetic influence through longitudinal clinical trajectories. These trajectories capture how patients accumulate acute and chronic diagnoses over time, and recent work suggests that sequential disease progression modeling can improve future risk estimation (Shmatko et al., 2025). Yet existing trajectory models often treat ICD-10 codes as categorical identifiers, leaving this semantic signal largely underused (Rasmy et al., 2021; Kim et al., 2025). This motivates a language-aware view of clinical trajectories, where pretrained language models can encode ICD-10 descriptions and sequence models can compose them with temporal context into patient-level phenotype representations.

Inspired by the semantic structure of disease language, we ask two questions: **RQ1:** Can language- and time-aware clinical trajectories provide semantic supervision for learning PRS-derived genetic representations? **RQ2:** Can the resulting pretrained genetic representations improve downstream disease risk prediction?

To answer these questions, we propose **Contrastive Phenotype–Genotype Pretraining (CPGP)**, a dual-encoder contrastive framework that aligns standardized PRS profiles with longitudinal ICD-10 histories in a shared embedding space. On the phenotype side, ICD-10 events are encoded as language- and time-aware tokens by combining pretrained text embeddings of medical descriptions with temporal embeddings for age-at-diagnosis and inter-event gaps; a bidirectional GRU (Cho et al., 2014; Schuster and Paliwal, 1997) with attention pooling then produces patient-level trajectory embeddings. The genetic branch projects pre-computed PRS profiles into the same latent

space, and the model is trained with a symmetric contrastive objective over matched and mismatched participant pairs. Our contributions are threefold:

- We introduce CPGP, a contrastive pretraining framework (Figure 1) that uses language- and time-aware ICD-10 trajectories as semantic supervision for learning transferable PRS-derived genetic representations.
- We show that CPGP learns a shared genotype–phenotype space through zero-shot retrieval and retrieval ablations against raw PRS and alternative ICD text encoders.
- We demonstrate the downstream utility of pretrained CPGP genetic embeddings for disease risk prediction across 30 endpoints against PRS baselines, and include a proof-of-concept agentic 10-year risk-estimation study showing that CPGP-retrieved patient neighborhoods can provide useful evidence under label-limited prediction.

2 Method

2.1 Data

We train and evaluate CPGP on UK Biobank participants with paired genetic risk profiles and longitudinal ICD-10 diagnosis histories. For participant i, the clinical trajectory is represented as a time-ordered sequence $\mathcal{C}_i = \{(c_{i,t}, a_{i,t})\}_{t=1}^{T_i}$, where $c_{i,t}$ denotes an ICD-10 diagnosis code and $a_{i,t}$ denotes age at diagnosis. Trajectories are constructed from first-occurrence ICD-10 records. Participants are split into training, validation, and test sets using a 70/15/15 split, yielding 341,399, 73,157, and 73,157 participants, respectively.

Each participant is also represented by a genetic risk vector $\mathbf{g}_i \in \mathbb{R}^{1153}$, consisting of 1,153 pre-computed PRS features derived from PGS Catalog (Lambert et al., 2021) sources. To reduce potential cohort leakage, we exclude PRS models developed on UK Biobank. PRS values are z-scored using training-set statistics only, and the same normalization parameters are applied to validation and test participants. For further preprocessing, see Appendix A.

2.2 Language-Guided Clinical History Encoder

The phenotype encoder maps each diagnosis trajectory $\mathcal{C}_i$ into a patient-level clinical embedding. For each ICD-10 event $c_{i,t}$, we construct a language- and time-aware token:

$$\mathbf{x}_{i,t} = f_{\text{text}}(e(c_{i,t})) + \mathbf{e}^{\text{age}}_{b(a_{i,t})} + \mathbf{e}^{\text{gap}}_{b(\Delta_{i,t})},$$

where $e(c_{i,t})$ is the frozen Qwen3-Embedding-4B (Zhang et al., 2025) representation of the ICD-10 textual description, f_{text} projects it from 2560 dimensions to $d = 128$, and $b(\cdot)$ maps age-at-diagnosis and inter-event gaps to discrete bins. Age is discretized into 5-year bins from 0 to 85+, and inter-event gaps are discretized into 2-year bins from 0 to 20+ years.

The resulting event sequence is encoded by a single-layer bidirectional GRU with packed sequences. Masked attention pooling aggregates non-padded hidden states into a patient-level trajectory representation, followed by projection and ℓ_2 normalization to obtain $\mathbf{z}_i^c \in \mathbb{R}^{128}$.

2.3 Genetic Encoder

The genetic encoder maps each standardized PRS vector $\mathbf{g}_i \in \mathbb{R}^{1153}$ to the shared 128-dimensional space using a two-layer MLP with LayerNorm and ℓ_2 normalization, producing $\mathbf{z}_i^g \in \mathbb{R}^{128}$.

2.4 Contrastive Language–Genotype Pretraining

CPGP learns a shared genotype–phenotype embedding space by aligning each participant's PRS-derived genetic representation with their language-encoded clinical trajectory. For a minibatch of N participants, we compute cross-modal similarities

$$s_{ij} = (\mathbf{z}_i^g)^\top \mathbf{z}_j^c / \tau,$$

where τ is a learnable temperature initialized to 0.07. We train CPGP with a symmetric InfoNCE objective (Oord et al., 2018):

$$\mathcal{L} = \frac{1}{2}\left[\text{CE}(\mathbf{S}, \mathbf{y}) + \text{CE}(\mathbf{S}^\top, \mathbf{y})\right],$$

where CE denotes cross-entropy loss, $\mathbf{S} = [s_{ij}]$ is the minibatch similarity matrix, and $\mathbf{y} = (1, \ldots, N)$ gives the matched-pair labels. The objective pulls matched genetic–clinical embeddings together and pushes mismatched participant pairs apart.

2.5 Retrieval-grounded Downstream Prediction

We use CPGP as an evidence-grounding backbone for few-shot 10-year disease risk estimation. For each target patient, an LLM agent receives PGS percentiles, lifestyle variables, and a target disease name. The agent can invoke two CPGP retrieval tools: PRS-to-PRS retrieval for genetically similar

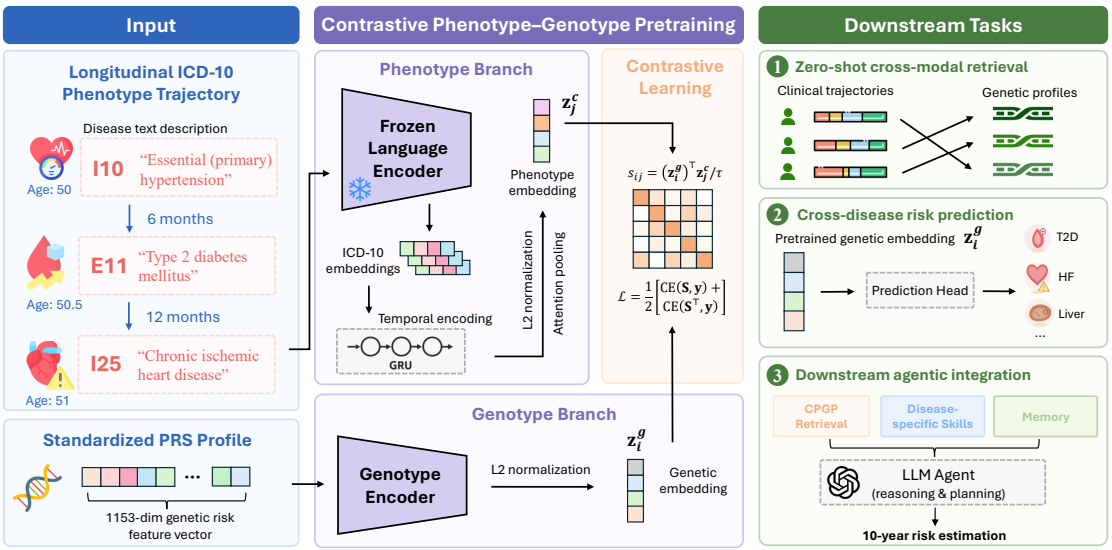

Figure 1: Overview of CPGP. Language- and time-aware ICD-10 trajectories supervise PRS-derived genetic representations through CLIP-style contrastive alignment, supporting zero-shot retrieval, cross-disease risk prediction, and agentic 10-year risk estimation.

patients and PRS-to-ICD retrieval for phenotype-relevant clinical trajectories. Guided by disease-specific skill files and constrained self-evolving memory, the agent converts genetic, lifestyle, and retrieval evidence into four structured evidence scores, which are passed to a lightweight logistic regression predictor. Implementation details are provided in Appendix D.

3 Experimental Results

3.1 Zero-shot Cross-modal Retrieval (RQ1)

Retrieval set up. This task evaluates whether CPGP transfers to unseen participants without training a downstream retrieval model. For each held-out participant, the frozen genetic encoder maps the PRS profile into the shared embedding space, and candidate clinical-history embeddings are ranked by cosine similarity. A query is correct if the matched clinical history appears among the top-k retrieved candidates. To assess scaling behavior, we train CPGP with increasing numbers of training participants while holding the validation and test splits fixed. We further compare against a raw PRS nearest-neighbor baseline, which ranks candidates by similarity in the original standardized PRS feature space without contrastive pretraining.

Retrieval results. As training size increases from 5K to the full training set of 341,399 participants, for held-out participants, top-1 retrieval accuracy improves from 0.017 to 0.098, while top-3 accuracy improves from 0.051 to 0.209 (Appendix B). Performance increases monotonically with no clear saturation, suggesting that contrastive phenotype–genotype alignment continues to benefit from additional biobank-scale supervision. At full scale, CPGP substantially outperforms raw PRS retrieval, which achieves only 0.013 at top-1 and 0.040 at top-3 accuracy. These results suggest that contrastive phenotype–genotype pretraining learns a shared embedding space that aligns PRS-derived genetic profiles with longitudinal clinical trajectories beyond raw PRS similarity. Additional retrieval ablations with different ICD-10 text encoders are reported in Appendix Table 3.

Endpoint	PGS	PRS	CPGP-L	CPGP-M
Macro AUROC	0.575	0.620	0.641	0.641
Macro AUPRC	0.105	0.126	0.137	0.138
Alzheimer's disease	0.658	0.652	0.675	0.673
Breast cancer	0.596	0.668	0.685	0.689
Heart failure	0.563	0.624	0.661	0.663
Ischaemic heart dis.	0.577	0.604	0.631	0.635
Liver disease	0.533	0.567	0.600	0.599
Sleep disorders	0.538	0.611	0.642	0.644
Stroke	0.519	0.547	0.586	0.583
Type 1 diabetes	0.621	0.760	0.804	0.805

Table 1: Cross-disease prediction using genetic information only. Macro rows report AUROC and AUPRC across 30 diseases. PGS refers to baseline (i), PRS refers to baseline (ii), and CPGP-L/M denote linear and MLP probes on frozen CPGP embeddings. Full 30-disease results are reported in Appendix C.

3.2 Cross-Disease Prediction with Frozen Genetic Embeddings (RQ1, RQ2)

PRS baselines. We next evaluate whether CPGP learns genetic representations that transfer across disease endpoints. After contrastive pretraining, we freeze the genetic encoder and extract a 128-

Table 2: Performance comparison of different methods for predicting 10-year incident hypertension, diabetes, and glaucoma on the 1,000-patient test cohort. Results are averaged from five runs, using gpt-oss-120b as the LLM backbone. F1 is computed using the top predicted cases according to the prevalence of each disease.

Method	Setting	Hypertension (40%)			Diabetes (10%)			Glaucoma (2.5%)		
		AUPRC	AUROC	F1	AUPRC	AUROC	F1	AUPRC	AUROC	F1
LR	Supervised	0.519	0.634	0.514	0.220	0.685	0.234	0.030	0.524	0.056
MLP	Supervised	0.547	0.649	0.528	0.214	0.696	0.256	0.051	0.565	0.080
CoT	Zero-shot	0.527	0.664	0.561	0.508	0.801	0.492	0.032	0.551	0.064
Agent w/o retrieval	Zero-shot	0.593	0.724	0.603	0.469	0.857	0.530	0.050	**0.696**	0.040
Agent w/o tools & skills	Few-shot	0.609	0.744	0.615	0.499	0.888	0.530	0.074	0.613	0.080
Agent w/o memory	Few-shot	0.617	0.747	0.615	0.556	0.876	0.570	0.046	0.683	0.040
Agent full	Few-shot	**0.637**	**0.756**	**0.622**	**0.615**	**0.893**	**0.612**	**0.091**	0.680	**0.120**

dimensional embedding for each participant. For each disease, we train a lightweight classifier on the frozen genetic embeddings and compare against two PRS-based baselines: (i) the mean performance of disease-related raw PGS Catalog scores and (ii) a linear classifier trained on all 1,153 standardized PRS features.

CPGP is evaluated with both a linear probe and a two-layer MLP probe. AUROC is used as the primary metric, and AUPRC as a secondary metric to account for class imbalance.

Cross-disease results. Across 30 disease prediction tasks, CPGP improves over both PRS-based baselines in macro AUROC and AUPRC. Against the PRS linear baseline trained on the same 1,153 standardized PRS features, frozen CPGP embeddings improve macro AUROC from 0.620 to 0.641 and macro AUPRC from 0.126 to 0.138. CPGP-L improves over PRS linear in all 30 diseases, with large gains for type 1 diabetes, heart failure, stroke, sleep disorders, and liver disease (Table 1). Because all predictors use genetic information only, without demographics, lifestyle, or longitudinal clinical covariates, the key result is the consistent improvement over PRS baselines using the same genetic input rather than the absolute AUROC magnitude. These results indicate that CPGP learns transferable PRS-derived genetic representations across clinically diverse endpoints.

3.3 Evaluating CPGP Utility in Agentic 10-Year Risk Estimation (RQ2)

We evaluate whether CPGP provides useful retrieval-based evidence in a modern agentic disease risk prediction setting. We select hypertension (40%), diabetes (10%), and glaucoma (2.5%) from UK Biobank to represent high-, moderate-, and low-prevalence disease prediction settings. We construct 1,000-patient training and test cohorts annotated for all three diseases. The training cohort is used for supervised baselines and self-evolving memory, while the test cohort is reserved for final evaluation. As shown in Table 2, the full agent improves over supervised LR and MLP baselines across the three diseases. Compared with the variant without CPGP retrieval, the full agent achieves a 20.8% average relative AUPRC improvement, suggesting that CPGP-retrieved genetic and phenotypic neighborhoods provide useful evidence for agentic risk estimation. For glaucoma, the most imbalanced task, the full agent improves over random guessing by 3.64× in AUPRC and 4.8× in F1, indicating potential utility for low-prevalence, label-limited diseases. Further setup details, self-evolving memory analysis, and LLM-backbone comparisons are provided in Appendices E, F, and G.

4 Conclusions

Our work introduces CPGP, a CLIP-style phenotype–genotype pretraining framework that uses longitudinal ICD-10 clinical trajectories as semantic supervision for transferable PRS-derived genetic representation learning. CPGP organizes inherited risk within a phenotype-aware embedding space rather than treating PRS features as isolated disease-specific predictors. Experiments on UK Biobank show that this shared space supports zero-shot cross-modal retrieval, improves genetic-only cross-disease prediction over PRS baselines, and provides useful retrieval evidence for label-limited 10-year disease risk estimation. These results highlight the value of clinical language trajectories as supervision for genotype–phenotype representation learning and point toward more interpretable disease risk modeling systems.

Limitations

This study has several limitations. First, CPGP is trained and evaluated on UK Biobank participants, who are predominantly of European ancestry. Future work should incorporate additional biobanks and healthcare systems with more diverse ancestral and demographic backgrounds to assess the generalizability and equity of the learned representations.

Second, CPGP uses precomputed PGS Catalog scores rather than raw genotype variants. While this provides a compact and interpretable genetic input space, it also makes the learned representations dependent on the coverage, quality, and population portability of existing GWAS and PGS resources.

Third, our agentic 10-year risk estimation experiment is a proof-of-concept application of CPGP retrieval rather than a deployable clinical prediction system. Although it evaluates whether CPGP-retrieved patient neighborhoods can support LLM-based reasoning across three diseases, broader evaluation across more endpoints, independent cohorts, larger label-limited settings, and stronger calibration protocols is needed.

Finally, CPGP is not intended for direct clinical deployment in its current form. The present study focuses on representation learning, cross-modal retrieval, and downstream risk prediction experiments. Before clinical use, the framework would require prospective validation, calibration, fairness analysis, uncertainty quantification, and careful assessment of real-world harms. Our results should therefore be interpreted as evidence for the representation-learning value of language- and time-aware clinical trajectories, rather than as clinically actionable risk predictions.

Ethics and Data Use

The research was conducted using data from the UK Biobank Resource under application number 34031. The UKB study has approval from the North West Multicenter Research Ethics Committee as a Research Tissue Bank approval.

References

Abdel Abdellaoui, Loic Yengo, Karin JH Verweij, and Peter M Visscher. 2023. 15 years of gwas discovery: realizing the promise. *The American Journal of Human Genetics*, 110(2):179–194.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, and 1 others. 2018. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209.

Christopher C Chang, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015.

Kyunghyun Cho, Bart Van Merriënboer, Çağlar Gulçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1724–1734.

Gemma Team, Google DeepMind. 2026. Gemma 4: Frontier multimodal intelligence on device. Accessed: May 25, 2026.

Minkyoung Kim, Yunha Kim, Hee Jun Kang, Hyeram Seo, Heejung Choi, JiYe Han, Gaeun Kee, Soyoung Ko, HyoJe Jung, Byeolhee Kim, and 1 others. 2025. Leveraging bert for embedding icd codes from large scale cardiovascular emr data to understand patient diagnostic patterns. *BMC Medical Informatics and Decision Making*, 25(1):300.

Iftikhar J Kullo. 2025. Clinical use of polygenic risk scores: current status, barriers and future directions. *Nature Reviews Genetics*, pages 1–18.

Samuel A Lambert, Laurent Gil, Simon Jupp, Scott C Ritchie, Yu Xu, Annalisa Buniello, Aoife McMahon, Gad Abraham, Michael Chapman, Helen Parkinson, and 1 others. 2021. The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nature genetics*, 53(4):420–425.

Cathryn M Lewis and Evangelos Vassos. 2020. Polygenic risk scores: from research tools to clinical instruments. *Genome medicine*, 12(1):44.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Qwen Team. 2026. Qwen3.6-35B-A3B: Agentic coding power, now open to all.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. 2025. Learning the natural history of human disease with generative transformers. *Nature*, 647(8088):248–256.

Barbara E Stranger, Eli A Stahl, and Towfique Raj. 2011. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics*, 187(2):367–383.

Ali Torkamani, Nathan E Wineinger, and Eric J Topol. 2018. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics*, 19(9):581–590.

Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina De Vries, Yukinori Okada, Alicia R Martin, Hilary C Martin, Tuuli Lappalainen, and Danielle Posthuma. 2021. Genome-wide association studies. *Nature reviews methods primers*, 1(1):59.

Ying Wang, Kristin Tsuo, Masahiro Kanai, Benjamin M Neale, and Alicia R Martin. 2022. Challenges and opportunities for developing more generalizable polygenic risk scores. *Annual review of biomedical data science*, 5(1):293–320.

Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. 2014. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Research*, 42(D1):D1001–D1006.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Genetic Features Preprocessing

We computed polygenic score features from the PGS Catalog. After scoring and merging catalog entries, 2,412 PGS traits had non-missing values in the scored UK Biobank cohort. To reduce cohort leakage, we used PGS Catalog (https://www.pgscatalog.org/) metadata to exclude scores whose development cohorts included UK Biobank or UKB-PPP samples. This removed 1,259 scores and retained 1,153 non-UKB-development PGS features.

Participant-level scores were computed using PLINK 1.9 `-score` (Chang et al., 2015), which applies the PGS effect weights to genotype dosages and sums dosage-weight products over variants shared between the scoring file and UK Biobank genotype data. The retained scores were concatenated into a 1,153-dimensional genetic feature vector for each participant and z-scored using training-set statistics only.

B Retrieval Analysis

Scaling with training size. Figure 2 shows zero-shot genetic-to-clinical-history retrieval accuracy under different CPGP training sizes. Retrieval accuracy improves monotonically as the number of training participants increases, suggesting that contrastive phenotype–genotype alignment continues to benefit from additional biobank-scale supervision.

Retrieval ablation. Table 3 reports full-scale retrieval ablations. We compare raw PRS nearest-neighbor retrieval with CPGP variants using different ICD-10 text encoders. This tests whether CPGP learns genotype–phenotype alignment beyond raw PRS similarity and whether retrieval is robust across clinical language encoders.

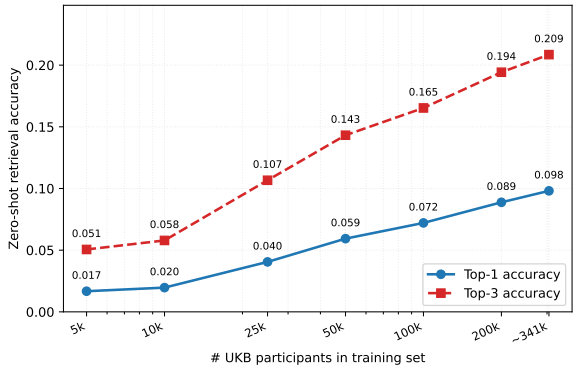

Figure 2: Zero-shot retrieval performance under different training sizes.

Retrieval model	Top-1 (%)	Top-3 (%)
Raw PRS retrieval	1.31	4.02
CPGP w/ SapBERT	9.35	20.15
CPGP w/ Qwen	9.81	20.85

Table 3: Zero-shot retrieval ablation at the full training scale. Raw PRS retrieval ranks candidates by similarity in the original standardized PRS feature space. SapBERT (Liu et al., 2021) and Qwen variants use different pretrained text encoders for ICD-10 descriptions while keeping the CPGP training objective unchanged.

C Full Genetic-Only Cross-Disease Prediction Results

See Table 4 for the full results.

D Details of the Agentic Reasoning Framework

D.1 Agent Workflow

We develop an agentic reasoning framework for disease risk assessment using CPGP and LLMs. For each target patient, the input consists of 1,153 PGS percentiles, a lifestyle profile, and the target disease name, and the output is a disease-specific risk estimate together with intermediate evidence scores. The agent first forms patient-level evidence from two sources: genetic risk evidence derived from disease-relevant PGS traits, and lifestyle evidence derived from metabolic and behavioral risk factors. It then follows a conditional reasoning workflow to determine whether additional retrieval is needed. If the genetic and lifestyle evidence provide strong and concordant risk signals, the agent can proceed without retrieval. If the evidence is weak, inconsistent, or incomplete, the agent selectively invokes retrieval tools to obtain cross-modal support from the CPGP space.

D.2 Disease-Specific Skills

To adapt the agent to different diseases, we design a lightweight disease-specific skill module. Each skill summarizes prior knowledge that is available before inference, including disease-relevant genetic traits, clinically meaningful lifestyle or contextual risk factors, and phenotype patterns that should be considered when interpreting retrieved patients. The skill is written as a human-readable Markdown file and is loaded by the agent at inference time. It is not optimized on the test set and does not contain patient labels from the evaluation cohort.

For each disease, the skill specifies a small set of relevant PGS traits selected from the available PGS metadata. For example, the diabetes mellitus skill identifies the PGS traits with the highest reported hazard ratios or odds ratios among diabetes and diabetes-related metabolic traits. A hazard ratio measures relative event risk over time, whereas an odds ratio measures the relative odds of disease association. These statistics are derived from the PGS Catalog. In addition, each skill lists lifestyle or clinical-context factors that are biologically or clinically relevant to the disease, such as body mass index, smoking, blood pressure, or age, depending on the disease. Details are shown in Figures 3, 4, and 5.

D.3 Retrieval and Prediction Tools

The agent is equipped with three external tools, two of which are built directly on the CPGP embedding space. The first tool performs PRS-to-PRS retrieval to identify genetically similar patients using PRS embeddings learned by CPGP. This allows the agent to examine whether the target disease is enriched among patients with similar genome-wide risk profiles. The second tool performs PRS-to-ICD retrieval by measuring similarity between the target patient's PRS representation and ICD embeddings in the shared CPGP space. This cross-modal retrieval provides phenotype-level evidence from other patients' clinical trajectories, enabling the agent to connect genetic risk profiles with disease-relevant diagnostic patterns.

After retrieval, the agent assigns four evidence scores: target PGS evidence, target lifestyle evidence, PRS-to-PRS retrieval evidence, and PRS-to-ICD retrieval evidence. Each score is on a 1 to 10 scale. Retrieval scores are generated only when the corresponding CPGP retrieval pathway is invoked or available. The third tool is a lightweight logis-

Disease	ICD-10	Cases	Prev.	PGS mean	PRS linear	CPGP-L	CPGP-M
Alzheimer's disease	G30, F00	581	1.3%	0.658	0.652	0.675	0.673
Anaemia	D50–D64	4635	10.6%	–	0.548	0.574	0.573
Anxiety disorders	F41	3014	6.7%	–	0.584	0.592	0.591
Arrhythmias excl. AF	I46, I47, I49	1621	3.6%	–	0.571	0.608	0.608
Asthma	J45	3093	7.3%	0.600	0.643	0.661	0.661
Atrial fibrillation	I48	4098	9.3%	0.622	0.675	0.688	0.689
Breast cancer	C50	1105	2.5%	0.596	0.668	0.685	0.689
Cardiovascular disease	I00–I99	14116	45.6%	0.500	0.599	0.616	0.616
Chronic kidney disease	N18	3288	7.3%	0.543	0.602	0.622	0.623
Colorectal cancer	C18–C20	822	1.8%	0.586	0.576	0.578	0.569
COPD	J44	2308	5.2%	0.568	0.651	0.662	0.663
Dementia	F00–F03, G30, G31	1428	3.1%	0.574	0.634	0.642	0.643
Diabetes mellitus	E10–E14	3686	8.6%	0.651	0.729	0.742	0.744
Essential hypertension	I10	11757	32.0%	0.637	0.657	0.669	0.670
Heart failure	I50	2218	4.9%	0.563	0.624	0.661	0.663
Ischaemic heart disease	I20–I25	4210	10.1%	0.577	0.604	0.631	0.635
Liver disease	K70–K77	1877	4.2%	0.533	0.567	0.600	0.599
Lung cancer	C34	633	1.4%	0.557	0.581	0.605	0.604
Major depressive disorder	F32, F33	3103	7.0%	0.560	0.597	0.607	0.605
Myocardial infarction	I21, I22	1419	3.2%	0.599	0.658	0.678	0.682
Obesity	E66	4484	10.1%	0.618	0.701	0.708	0.709
Osteoarthritis	M15–M19	8421	20.8%	0.534	0.554	0.581	0.580
Osteoporosis	M80, M81	2353	5.3%	0.475	0.559	0.575	0.576
Peripheral artery disease	I70–I79	2054	4.6%	0.501	0.575	0.602	0.601
Renal failure	N17–N19	5402	12.0%	–	0.584	0.607	0.608
Rheumatoid arthritis	M05, M06	806	1.8%	0.564	0.559	0.572	0.571
Sleep disorders	G47	1128	2.5%	0.538	0.611	0.642	0.644
Stroke	I60–I69	2633	5.9%	0.519	0.547	0.586	0.583
Type 1 diabetes	E10	278	0.6%	0.621	0.760	0.804	0.805
Type 2 diabetes	E11	3604	8.3%	0.658	0.739	0.749	0.752
Macro AUROC	–	–	–	**0.575**	**0.620**	**0.641**	**0.641**
Macro AUPRC	–	–	–	**0.105**	**0.126**	**0.137**	**0.138**

Table 4: Full cross-disease prediction results using genetic information only. Disease rows report AUROC. Macro rows report average AUROC and AUPRC across 30 disease endpoints, except PGS mean, which is averaged over the 26 available disease-related PGS scores. PGS mean denotes the mean performance of available disease-related raw PGS Catalog scores; PRS linear denotes a linear classifier trained on all 1,153 standardized PRS features; CPGP-L and CPGP-M denote linear and two-layer MLP probes on frozen CPGP genetic embeddings. A dash indicates that no directly matched disease-related PGS score was available under our PGS selection criteria.

tic regression predictor, which takes the four evidence scores as input and produces the final disease-specific risk estimate. The predictor is trained on the training set.

D.4 Self-Evolving Memory

We introduce a self-evolving memory mechanism to improve agentic reasoning without updating the LLM backbone. For each disease, the training cohort contains 1,000 patients and follows the same disease prevalence as the corresponding test cohort. None of these patients are used for CPGP pretraining. The training cohort is divided into five batches of 200 patients. To avoid reinforcing misleading prevalence patterns, each batch is constructed to include a balanced set of positive cases, so that the agent receives sufficient feedback on true disease cases throughout the memory construction process.

After processing each batch, the agent compares its predictions with the ground-truth labels and writes a compact Markdown memory file. Each memory file summarizes the scope of the batch, false-positive experience, true-positive rescue experience, PRS-to-PRS versus PRS-to-ICD conflict experience, batch-level lessons, lifestyle evidence patterns, evidence-weighting experience, and true-positive harm checks. These entries are designed to record not only common error patterns but also concrete correction strategies, such as when lifestyle evidence should down-weight weak genetic signals, when retrieval evidence can rescue borderline positive cases, or when conflicting PRS-to-PRS and PRS-to-ICD evidence should be treated cautiously.

At test time, the agent first performs the initial evidence assessment and assigns four evidence scores: target PGS evidence, target lifestyle evidence, PRS-to-PRS retrieval evidence, and PRS-

hypertension.md

Purpose. The hypertension skill provides disease-specific guidance for identifying relevant genetic and lifestyle evidence. It includes lifestyle and clinical risk factors, direct hypertension-related PGS traits, and hypertension-related cardiometabolic PGS traits. The selected PGS traits correspond to hypertension or hypertension-related cardiometabolic traits with high reported hazard ratios or odds ratios in the PGS Catalog.

Lifestyle and clinical factors.

- **Systolic blood pressure:** higher risk if systolic blood pressure $\geq$ 140 mmHg; moderate risk if systolic blood pressure is between 130 and 140 mmHg.
- **BMI:** higher BMI supports elevated hypertension risk.
- **Smoking:** current smoking supports cardiovascular risk context.
- **Physical activity:** low MET-minutes supports elevated risk, whereas high activity provides protective context.
- **Age:** older age supports higher baseline hypertension risk.
- **Creatinine:** higher creatinine supports kidney-hypertension risk context.
- **LDL cholesterol:** high LDL cholesterol supports cardiovascular risk context.

Direct hypertension PGS traits.

PGS ID	Reported trait
PGS002296	Hypertension
PGS004191	Hypertension
PGS004192	Hypertension
PGS004193	Hypertension
PGS004194	Hypertension
PGS004195	Hypertension
PGS005120	Systolic blood pressure
PGS000301	Systolic blood pressure
PGS002275	Systolic blood pressure
PGS002734	Systolic blood pressure

Hypertension-related cardiometabolic PGS traits.

PGS ID	Reported trait
PGS000822	Urinary albumin to creatinine ratio
PGS003400	Obesity
PGS005168	Atrial fibrillation
PGS004928	Metabolic syndrome
PGS003404	LDL cholesterol

Figure 3: Skill markdown used by the agent for hypertension.

diabetes.md

Purpose. The diabetes mellitus skill provides disease-specific guidance for identifying relevant genetic and lifestyle evidence. It includes lifestyle risk factors, direct diabetes-related PGS traits, and diabetes-related metabolic PGS traits. The selected PGS traits correspond to diabetes or diabetes-related metabolic traits with high reported hazard ratios or odds ratios in the PGS Catalog.

Lifestyle factors.

- **BMI:** higher risk if BMI $\geq$ 30; moderate risk if BMI is between 25 and 30.
- **Smoking:** current smoking supports elevated risk; former smoking provides weaker support.
- **Physical activity:** low MET-minutes supports elevated risk, whereas high activity provides protective context.
- **Age:** older age supports higher baseline risk.
- **Socioeconomic deprivation:** higher Townsend deprivation provides additional risk context.

Direct diabetes PGS traits.

PGS ID	Reported trait
PGS003353	Type 2 diabetes (T2D)
PGS005110	Type 2 diabetes (T2D)
PGS004874	Type 1 diabetes (T1D)
PGS000023	Type 1 diabetes (T1D)
PGS002308	Type 2 diabetes (T2D)
PGS005246	Type 2 diabetes (T2D)
PGS000729	Type 2 diabetes (T2D)
PGS003445	Type 2 diabetes (T2D)
PGS003443	Type 2 diabetes (T2D)
PGS003444	Type 2 diabetes (T2D)

Diabetes-related metabolic PGS traits.

PGS ID	Reported trait
PGS003336	Fasting plasma glucose
PGS000839	Two-hour glucose during OGTT
PGS000838	Fasting glucose
PGS003337	HbA1c
PGS000871	Insulin secretion

Figure 4: Skill markdown used by the agent for diabetes.

to-ICD retrieval evidence. Only after these initial scores are assigned does the agent read the self-evolving memory. Based on the memory, the agent may adjust at most two of the four evidence scores. Each adjustment is constrained to the range of $[-3, +3]$, and the final calibrated evidence scores are constrained to remain within the valid score range of 1–10. This constrained calibration allows the memory to correct recurring reasoning errors while preserving the original evidence structure

glaucoma.md

Purpose. The glaucoma skill provides disease-specific guidance for identifying relevant genetic and lifestyle evidence. It includes lifestyle and clinical risk factors, direct glaucoma-related PGS traits, and glaucoma-related ocular PGS traits. The selected PGS traits correspond to glaucoma or glaucoma-related ocular traits with high reported hazard ratios or odds ratios in the PGS Catalog.

Lifestyle and clinical factors.

- **Age:** older age supports higher baseline glaucoma risk.
- **Diabetes context:** diabetes or high HbA1c can support ocular complication context.
- **Hypertension context:** elevated systolic blood pressure or hypertension supports vascular eye-risk context.
- **Smoking:** current smoking supports vascular risk context.
- **BMI:** high BMI provides weak contextual support only.

Direct glaucoma PGS traits.

PGS ID	Reported trait
PGS000350	Primary open-angle glaucoma
PGS000137	Glaucoma
PGS000764	Primary open-angle glaucoma
PGS001792	Primary open-angle glaucoma
PGS002741	Primary open-angle glaucoma

Glaucoma-related ocular PGS traits.

PGS ID	Reported trait
PGS000124	Intraocular pressure
PGS000879	Intraocular pressure
PGS000862	Diabetic retinopathy
PGS004606	Age-related macular degeneration
PGS004607	Age-related macular degeneration

Figure 5: Skill markdown used by the agent for glaucoma.

and preventing the memory from overriding the primary patient-level and retrieval-based evidence.

D.5 Prompt Templates

Agent Prompts

```
# Agent Prompts

## Prompt 1: Target Evidence Scoring Before Retrieval
      Planning

This prompt is called after the disease skill is parsed
      and after the target patient's disease-relevant PGS
      and lifestyle evidence are loaded. It outputs
      exactly two initial scores that are passed to
      retrieval planning.

### System Prompt

```text
You are Step 2 of a <disease> research demo agent. Read
      the disease skill, the target patient's disease-
      relevant PGS percentiles, and the target patient's
      lifestyle/clinical covariates. Assign exactly two
      initial evidence scores before retrieval planning:
      target_prs_score and target_lifestyle_score. Each
      score must be an integer from 1 to 10, where 1 means
       minimal evidence, 4 means weak/moderate evidence, 7
       means strong evidence, and 10 means very strong
      evidence. The scores should reflect evidence for the
       target disease only. Do not call retrieval tools,
      do not use labels, do not output a final diagnosis,
      and do not make memory-based corrections. Return
      JSON only.
```

User Payload

```json
{
  "task": "Score target PRS and lifestyle evidence for <
      disease> before retrieval planning.",
  "target": {
    "patient_id": "<patient_id>",
    "disease": "<disease>"
  },
  "skill": "<skill_text>",
  "target_evidence": {
    "genetic": {
      "overall_level": "<rule-derived level before LLM
      scoring>",
      "direct_<disease>": "<summary of disease-direct PGS
      percentiles>",
      "secondary_supporting": "<summary of secondary/
      supporting PGS percentiles>",
      "generic_top_pgs": "<summary of top generic PGS if
      no skill is used>",
      "all_skill_pgs": "<all skill-selected PGS records
      sorted by percentile>"
    },
    "lifestyle": {
      "overall_level": "<rule-derived level before LLM
      scoring>",
      "risk_points": "<rule-derived lifestyle risk points
      >",
      "factors": "<interpreted lifestyle/clinical factors
      >",
      "raw_lifestyle": "<raw lifestyle/clinical covariates
      >"
    }
  },
  "required_json_schema": {
    "target_prs_score": "integer 1-10",
    "target_lifestyle_score": "integer 1-10",
    "target_prs_rationale": "brief explanation based on
      skill-listed PGS percentiles",
    "target_lifestyle_rationale": "brief explanation based
       on lifestyle/clinical covariates"
  }
}
```

Expected JSON Response

```json
{
  "target_prs_score": 7,
  "target_lifestyle_score": 6,
  "target_prs_rationale": "The skill-listed disease PGS
      are concentrated at high percentiles.",
  "target_lifestyle_rationale": "The lifestyle and
      clinical covariates show moderate disease-relevant
      support."
}
```

The runner normalizes both scores to integers in `[1,
      10]`. These scores are converted into planning
      levels:

```text
1-3  -> low
4-6  -> moderate
```
```

7-10 -> high
```

## Prompt 2: Retrieval Decision

This prompt receives the LLM-scored target PRS and lifestyle evidence. It decides whether to call retrieval tools before downstream evidence integration.

### System Prompt

```text
You are Step 2/3 of a <disease> research demo agent. Read the skill, target PRS evidence, lifestyle evidence, and tool contracts. Decide whether the agent can diagnose immediately or should call retrieval tools. Return JSON only. Required keys: genetic_evidence_level, lifestyle_evidence_level, conflict, preliminary_diagnosis, selected_tools, rationale. selected_tools must be a list containing only prs_to_prs and/or prs_to_icd. Use prs_to_prs when genetic evidence is meaningful and genetically similar patients would help. Use prs_to_icd when phenotype support is needed or evidence is weak/conflicting. If both genetic and lifestyle evidence are high and concordant, selected_tools may be empty. This is a demo, not clinical medical advice.
```

User Payload

```json
{
  "task": "Step 2/3 decision for <disease> demo agent.",
  "target": {
    "patient_id": "<patient_id>",
    "disease": "<disease>"
  },
  "skill": "<skill_text>",
  "tool_contracts": "<tool_readme>",
  "target_evidence": {
    "genetic": {
      "overall_level": "<LLM-score-derived planning level>",
      "direct_<disease>": "<summary of disease-direct PGS percentiles>",
      "secondary_supporting": "<summary of secondary/supporting PGS percentiles>",
      "generic_top_pgs": "<summary of top generic PGS if no skill is used>",
      "all_skill_pgs": "<all skill-selected PGS records sorted by percentile>",
      "llm_target_evidence_score": {
        "score": "<target_prs_score>",
        "level": "<low|moderate|high>",
        "rule_score": "<deterministic PRS score before LLM scoring>",
        "rationale": "<target_prs_rationale>"
      }
    },
    "lifestyle": {
      "overall_level": "<LLM-score-derived planning level>",
      "risk_points": "<rule-derived lifestyle risk points>",
      "factors": "<interpreted lifestyle/clinical factors>",
      "raw_lifestyle": "<raw lifestyle/clinical covariates>",
      "llm_target_evidence_score": {
        "score": "<target_lifestyle_score>",
        "level": "<low|moderate|high>",
        "rule_score": "<deterministic lifestyle score before LLM scoring>",
        "rationale": "<target_lifestyle_rationale>"
      }
    }
  },
  "rule_baseline": {
    "preliminary_judgment": {
      "genetic_level": "<low|moderate|high|unavailable>",
      "lifestyle_level": "<low|moderate|high|unavailable>",
      "conflict": "<boolean>",
      "action": "<direct_positive_without_retrieval|needs_step3_retrieval>",
      "preliminary_diagnosis": "<positive|defer>",
      "reason": "<rule explanation>"
    },
    "selected_tools": [
      "prs_to_prs",
      "prs_to_icd"
    ]
  },
  "return_json_schema": {
    "genetic_evidence_level": "high|moderate|low|unavailable",
    "lifestyle_evidence_level": "high|moderate|low|unavailable",
    "conflict": "boolean",
    "preliminary_diagnosis": "positive|negative|defer",
    "selected_tools": [
      "prs_to_prs",
      "prs_to_icd"
    ],
    "rationale": "short explanation"
  }
}
```

Expected JSON Response

```json
{
  "genetic_evidence_level": "high",
  "lifestyle_evidence_level": "moderate",
  "conflict": false,
  "preliminary_diagnosis": "defer",
  "selected_tools": [
    "prs_to_prs",
    "prs_to_icd"
  ],
  "rationale": "The target PRS evidence is meaningful but lifestyle support is not jointly high, so retrieval evidence is useful before final scoring."
}
```

If `--force-tools both`, `--force-tools prs_to_prs`, `--force-tools prs_to_icd`, or `--force-tools none` is used, this prompt is bypassed and the forced tool selection is used.

Prompt 3: Memory-Guided Score Review

This prompt is called after target evidence and retrieval evidence are converted into four evidence scores. It can apply bounded memory-based calibration, but it cannot output a final diagnosis.

System Prompt

```text
You are a bounded <disease> evidence calibration layer. You receive memory-free base evidence scores and a disease experience memory. The memory is experience, not authority. You must not make a final diagnosis. You may only make bounded corrections to the base scores: adjust at most two scores, and each changed score may move by any integer delta from -3 to +3. You are not required to change any score, not required to change two scores, and not required to use the full three-point movement; choose 0, +/-1, +/-2, or +/-3 only when justified. If memory suggests both false-positive caution and true-positive rescue, prefer preserving the base score. Do not turn all evidence sources low unless the base evidence is already low. Return JSON only.
```

User Payload

```json
{
  "task": "Review memory-free base evidence scores and apply only bounded memory calibration. Do not output positive/negative diagnosis.",
  "disease": "<disease>",
  "calibration_bounds": {
    "memory_role": "experience only; it cannot override observed evidence by itself",
```
```

```
    "max_changed_scores": 2,
    "max_absolute_delta_per_score": 3,
    "allowed_delta_values": [
      -3,
      -2,
      -1,
      0,
      1,
      2,
      3
    ],
    "change_is_optional": true,
    "full_delta_is_optional": true,
    "allowed_adjustment_types": [
      "false_positive_caution",
      "true_positive_rescue",
      "conflict_resolution",
      "no_change"
    ],
    "preserve_base_when_uncertain": true
  },
  "score_scale": {
    "0": "available evidence is explicitly absent, non-
      supporting, or contradictory",
    "1": "minimal evidence",
    "4": "weak/moderate evidence",
    "7": "strong evidence",
    "10": "very strong evidence",
    "null": "evidence source unavailable or not triggered;
       do not impute"
  },
  "required_json_schema": {
    "calibration_answers": [
      {
        "question": "string",
        "answer": "short evidence-grounded answer"
      }
    ],
    "adjusted_scores": {
      "target_prs": "integer 1-10|null",
      "target_lifestyle": "integer 1-10|null",
      "prs_neighbor": "integer 1-10|null",
      "prs_to_icd": "integer 1-10|null"
    },
    "adjustment_rationale": {
      "target_prs": "short reason",
      "target_lifestyle": "short reason",
      "prs_neighbor": "short reason",
      "prs_to_icd": "short reason"
    },
    "changed_scores": [
      {
        "name": "target_prs|target_lifestyle|prs_neighbor|
      prs_to_icd",
        "base_score": "integer 0-10|null",
        "adjusted_score": "integer 0-10|null",
        "delta": "-3|-2|-1|0|1|2|3",
        "adjustment_type": "false_positive_caution|
      true_positive_rescue|conflict_resolution|no_change",
        "reason": "short reason grounded in memory and
      evidence"
      }
    ],
    "dominant_signal": "short phrase",
    "main_contradiction": "short phrase"
  },
  "disease_memory": "<memory_text>",
  "skill": "<skill_text>",
  "rule_evidence_scores": {
    "target_prs": {
      "available": "<boolean>",
      "score": "<integer 1-10|null>",
      "rule_score": "<integer 1-10|null, present if LLM
      target evidence scoring overrides the deterministic
      score>",
      "score_source": "<llm_target_evidence_scoring if
      overridden>",
      "llm_rationale": "<target PRS rationale if
      overridden>",
      "score_definition": "Ordinal 1-10 disease-relevant
      PRS evidence score; 1=minimal, 4=weak/moderate, 7=
      strong, 10=very strong.",
      "features": "<target PRS evidence features>",
      "rationale": "<rule or LLM-supported rationale>"
    },
    "target_lifestyle": {
      "available": "<boolean>",
      "score": "<integer 1-10|null>",
      "rule_score": "<integer 1-10|null, present if LLM
      target evidence scoring overrides the deterministic
      score>",
      "score_source": "<llm_target_evidence_scoring if
      overridden>",
      "llm_rationale": "<target lifestyle rationale if
      overridden>",
      "score_definition": "Ordinal 1-10 lifestyle/
      contextual risk evidence score; 1=minimal, 4=weak/
      moderate, 7=strong, 10=very strong.",
      "features": "<target lifestyle evidence features>",
      "rationale": "<rule or LLM-supported rationale>"
    },
    "prs_neighbor": {
      "available": "<boolean>",
      "score": "<integer 1-10|null>",
      "score_definition": "Ordinal 1-10 PRS-neighbor
      disease support score; 1=low, 4=weak/moderate, 7=
      strong, 10=very strong.",
      "features": "<PRS-to-PRS retrieval summary or
      unavailable reason>",
      "rationale": "<retrieval score rationale>"
    },
    "prs_to_icd": {
      "available": "<boolean>",
      "score": "<integer 1-10|null>",
      "score_definition": "Ordinal 1-10 PRS-to-ICD
      phenotype support score; 1=low, 4=weak/moderate, 7=
      strong, 10=very strong.",
      "features": "<PRS-to-ICD retrieval summary or
      unavailable reason>",
      "rationale": "<retrieval score rationale>"
    },
    "note": "Scores are ordinal evidence scores from 1 to
      10. Null means the evidence source was not observed
      or unavailable."
  },
  "integrated_evidence": {
    "target_prs_evidence": "<original target PRS evidence
      summary>",
    "target_lifestyle_evidence": "<original target
      lifestyle evidence summary>",
    "planning_target_prs_evidence": "<LLM-scored target
      PRS evidence used for planning>",
    "planning_target_lifestyle_evidence": "<LLM-scored
      target lifestyle evidence used for planning>",
    "prs_neighbor_evidence": "<PRS-to-PRS retrieval
      evidence summary or null>",
    "prs_to_icd_phenotype_evidence": "<PRS-to-ICD
      retrieval evidence summary or null>"
  }
}
```

Expected JSON Response

```json
{
  "calibration_answers": [
    {
      "question": "Which evidence source most strongly
      supports the base score?",
      "answer": "The strongest support comes from PRS-to-
      ICD phenotype evidence."
    }
  ],
  "adjusted_scores": {
    "target_prs": 7,
    "target_lifestyle": 6,
    "prs_neighbor": 5,
    "prs_to_icd": 8
  },
  "adjustment_rationale": {
    "target_prs": "No memory-based change.",
    "target_lifestyle": "No memory-based change.",
    "prs_neighbor": "Slightly reduced because memory flags
       weak PRS-neighbor support as noisy.",
    "prs_to_icd": "Preserved because phenotype retrieval
      is strongly disease-relevant."
  },
  "changed_scores": [
    {
      "name": "prs_neighbor",
      "base_score": 6,
```

```
      "adjusted_score": 5,
      "delta": -1,
      "adjustment_type": "false_positive_caution",
      "reason": "Memory indicates isolated PRS-neighbor
      support can overcall positives when phenotype
      evidence is weaker."
    }
  ],
  "dominant_signal": "PRS-to-ICD phenotype evidence",
  "main_contradiction": "PRS-neighbor evidence is weaker
      than phenotype retrieval"
}
```

The runner enforces the bounds after the LLM returns:

```text
at most two changed scores
each changed score delta must be between -3 and +3
available scores cannot be converted to null
```

Prompt 4: Legacy Final Diagnosis

This prompt is only used when `--diagnosis-mode legacy` is
       selected. It is not used in the current score-only
      v5 full-agent workflow.

System Prompt

```text
You are a research demo agent for <disease> assessment.
      Use only the provided evidence. Return concise JSON
      with keys: diagnosis (positive/negative), confidence,
       rationale, key_evidence, limitations. The final
      diagnosis must be binary: choose either positive or
      negative. Do not return uncertain. If disease
      experience memory is provided, use it as contextual
      experience for weighing evidence, not as a hard rule
       or a replacement for the observed patient evidence.
       This is not clinical medical advice.
```

User Payload

```json
{
  "task": "Integrate Step 1-4 evidence and produce Step 5
      final <disease> demo diagnosis.",
  "skill": "<skill_text>",
  "tool_contracts": "<tool_readme>",
  "disease_experience_memory": "<memory_text|null>",
  "evidence": {
    "target_prs_evidence": "<target PRS evidence summary
      >",
    "target_lifestyle_evidence": "<target lifestyle
      evidence summary>",
    "planning_target_prs_evidence": "<LLM-scored target
      PRS evidence used for planning>",
    "planning_target_lifestyle_evidence": "<LLM-scored
      target lifestyle evidence used for planning>",
    "prs_neighbor_evidence": "<PRS-to-PRS retrieval
      evidence summary or null>",
    "prs_to_icd_phenotype_evidence": "<PRS-to-ICD
      retrieval evidence summary or null>"
  }
}
```

Expected JSON Response

```json
{
  "diagnosis": "positive",
  "confidence": "moderate",
  "rationale": "Concise explanation integrating target
      evidence and retrieval evidence.",
  "key_evidence": [
    "Evidence item 1",
    "Evidence item 2"
  ],
  "limitations": [
    "This is a research demo and not clinical medical
      advice."
  ]
}
```
```

## E Agent Experiment Setup

We provide the experimental setup for the agentic disease risk assessment experiments to support reproducibility. The main LLM backbone is gpt-oss-120b (OpenAI, 2025). In the full agent setting, inference is performed with vLLM 0.21.0. We use a decoding temperature of 0.1 and keep the remaining decoding parameters at their default values. All agent experiments are run on 2× NVIDIA RTX PRO 6000 Blackwell GPUs with 96GB memory each.

For each disease, we construct a 1,000-patient training cohort and a 1,000-patient test cohort. The disease prevalence is matched between the training and test cohorts, with 40% for hypertension, 10% for type 2 diabetes, and 2.5% for glaucoma. These correspond to 400, 100, and 25 positive cases in each cohort, respectively. The training cohort is used to train the supervised baselines and to construct the batch-wise self-evolving memory for the agent. The test cohort is held out and used only for final evaluation.

We evaluate performance using AUROC, AUPRC, and F1@top-$K$. For each disease, $K$ is set to the number of positive cases in the test cohort, so that $K = 400$ for hypertension, $K = 100$ for type 2 diabetes, and $K = 25$ for glaucoma. All reported results are averaged over five random runs.

For baseline configurations, logistic regression (LR) and multi-layer perceptron (MLP) are implemented in PyTorch using each patient's 1,153 raw PGS values concatenated with available lifestyle/-clinical covariates. The chain-of-thoughts (CoT) baseline queries the LLM once per patient without disease-specific skills, memory, retrieval tools, or ICD evidence, using only top disease-agnostic PGS percentile records and available lifestyle/clinical covariates to directly output a binary diagnosis and disease probability in JSON format. See CoT's prompt below.

**CoT Zero-Shot Baseline Prompt**

```
System Prompt

```text
You are a zero-shot clinical genetics baseline. Use only
      the provided PRS percentiles and lifestyle profile.
      Think step by step, but keep the written reasoning
      brief. Return JSON only with keys: diagnosis,
```
```
```

```
      probability, confidence, reasoning. diagnosis must
      be exactly positive or negative. probability must be
       the probability that the patient has the target
      disease, from 0 to 1.
`´´´´

## User Payload Template

The user message is a JSON object generated by `
      build_prompt_patient(...)`:

´´´´
```

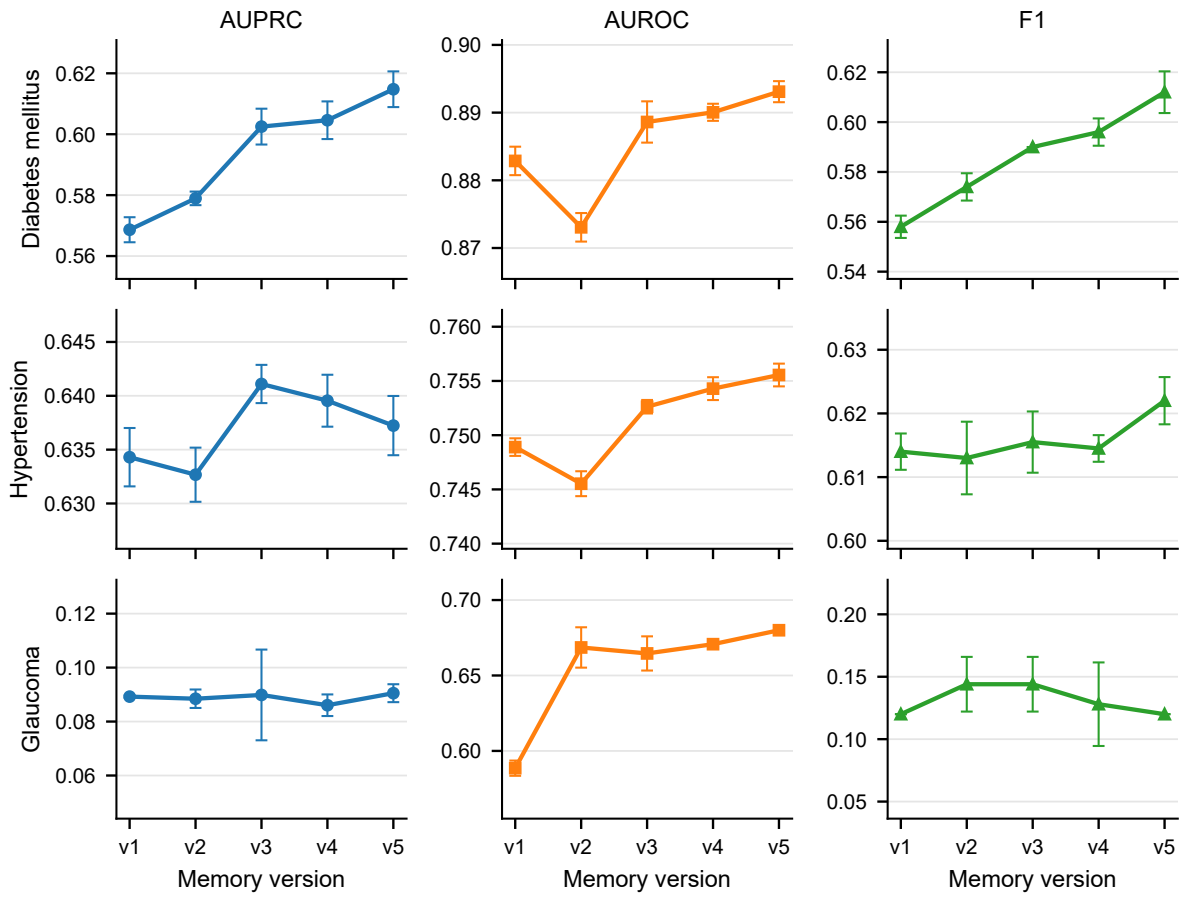

Figure 6: Performance across self-evolving memory versions for three disease prediction tasks. Versions v1–v5 denote sequential memory updates after each training batch. AUPRC, AUROC, and F1 are reported for diabetes mellitus, hypertension, and glaucoma; points show means and error bars show variability across 5 runs.

port; and **Qwen3.6-35B-A3B**[3] (Qwen Team, 2026), a Qwen open-weight model explicitly supporting vLLM, SGLang, and Transformers-based serving.

Across all three diseases, performance remains highly similar across LLM backends. For diabetes, the three models achieve AUROC around 0.89, AUPRC around 0.61, and F1@K around 0.60. For hypertension, the three models are nearly indistinguishable, with AUROC around 0.76, AUPRC around 0.64, and F1@K around 0.62. For glaucoma, all models remain constrained by the small number of positive cases, and F1@K is identical across backends. These results suggest that, once the LLM can reliably follow structured prompts and return valid JSON, the full agent's performance is driven primarily by the fixed evidence pipeline: disease skills, retrieval evidence, and memory-guided score calibration.

H Reproducibility

Code is available at `https://anonymous.4open.science/r/CPGP-0311/`.

[3] https://huggingface.co/Qwen/Qwen3.6-35B-A3B

Example Self-Evolving Memory (version 5) for Diabetes

Scope.

- Past experience suggests that memory guides calibration of four evidence streams—target PRS, lifestyle/biomarkers, PRS-to-PRS neighbor prevalence, and PRS-to-ICD retrieval—after a base score is produced.
- Past experience suggests that disease-proximal biomarkers such as diabetes-range HbA1c can dominate when genetic and phenotype signals conflict.
- Past experience suggests that occasional false-negative pressure warrants checking for distributed moderate signals before lowering risk.

False Positive Experience.

- High target PRS together with generic cardiometabolic lifestyle scores may create false-positive pressure when PRS-to-PRS neighbor prevalence and direct diabetes ICD counts are low; check for diabetes-specific ICD evidence or HbA1c $\geq 6.5\%$ before raising confidence.
- Broad metabolic ICD categories, such as obesity, hypertension, or kidney-related codes, without explicit diabetes codes can inflate risk; verify diabetes-range HbA1c or direct diabetes PGS before confirming.
- Strong PRS-to-PRS enrichment alone may be noisy when neighbors lack labeled diabetes cases; cross-check with target PRS percentile and diabetes-specific biomarkers before increasing the score.

True Positive Rescue Experience.

- Diabetes-range HbA1c, defined as $\geq 6.5\%$, or pre-diabetes HbA1c, defined as 5.7–6.4%, can rescue a true positive even when neighbor prevalence is low and lifestyle fields are missing.
- Moderate target PRS combined with diabetes-range HbA1c may preserve elevated risk despite low PRS-to-PRS and modest ICD counts.
- Several secondary signals, such as fasting-glucose PGS and hypertension-related lifestyle evidence, can collectively indicate genuine risk when no single source dominates.

PRS-to-PRS vs. PRS-to-ICD Conflict Experience.

- Direct diabetes ICD hits should be weighted higher than weak PRS-to-PRS enrichment when ICD lists contain E10–E14 codes.
- Strong PRS-to-PRS diabetes enrichment can support risk, but should be checked against target PRS percentile and diabetes-specific biomarkers before increasing confidence.
- When both PRS-to-PRS and PRS-to-ICD only show secondary metabolic codes, a small or no adjustment is usually safest.

Batch Lessons.

- **False-positive lesson:** broad cardiometabolic lifestyle or ICD codes without direct diabetes evidence repeatedly generate false-positive pressure; verify diabetes-specific ICD evidence or HbA1c before raising risk.
- **False-negative / true-positive rescue lesson:** requiring every evidence source to be strong can suppress true positives; distributed moderate signals across genetics, biomarkers, and phenotype retrieval can preserve elevated scores.

Lifestyle.

- Diabetes-range HbA1c provides disease-proximal evidence and can outweigh noisy cardiometabolic lifestyle factors such as smoking or activity level.
- Generic risk factors such as high blood pressure and obesity codes may support risk but can be misleading when genetic and ICD evidence are absent.
- Missing lifestyle fields should be treated as unknown rather than evidence against diabetes, and a single strong biomarker can rescue risk assessment.

Evidence Weight Experience.

- Direct diabetes PGS and ICD codes usually carry more weight than secondary glucose PGS or generic cardiometabolic traits.
- PRS-to-PRS enrichment is useful when neighbor prevalence includes labeled diabetes cases, but unstable when neighbors lack explicit labels.
- Lifestyle evidence should be integrated with genetic and phenotype signals, rather than used as the sole driver of calibration.

True Positive Harm Check.

- Downgrading risk solely because lifestyle fields are missing may harm true positives; before downgrading, check for diabetes-specific ICD hits, high direct PGS, or diabetes-range HbA1c.
- Requiring extreme PRS percentiles from many diabetes-related PGS before preserving risk may also harm true positives; moderate signals across multiple sources can collectively indicate risk.

Figure 7: The latest version (v5) of self-evolving memory generated after batch-wise feedback.

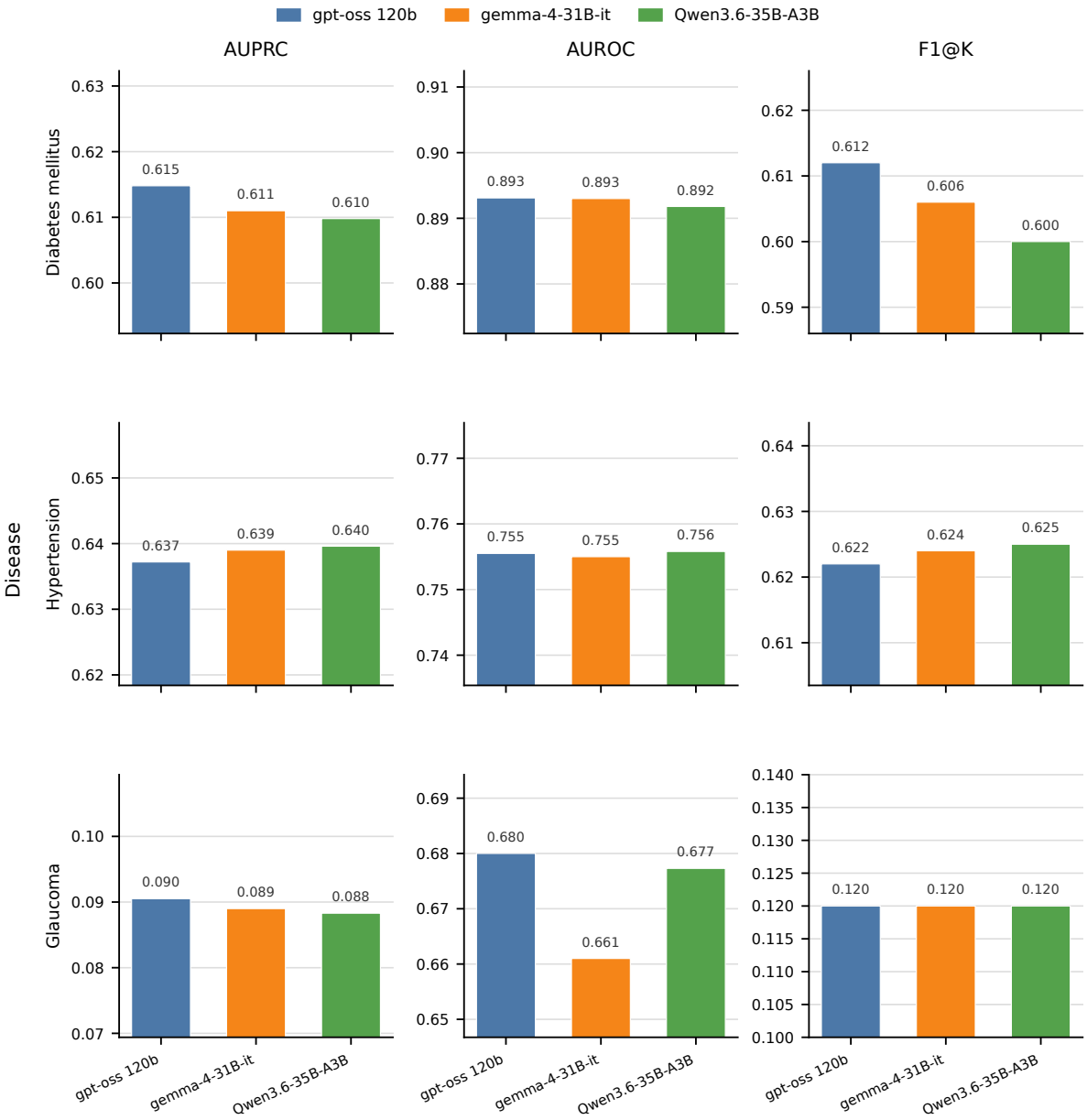

Figure 8: Performance comparison of the v5 full agent using different LLM backends across three diseases. Each subplot reports one evaluation metric for one disease on the same 1k held-out test set. Only the LLM backend was changed. F1 was computed using top-K prediction, where K equals the number of positive cases for each disease. Results are averaged from 5 runs.