nature medicine

Supplementary information

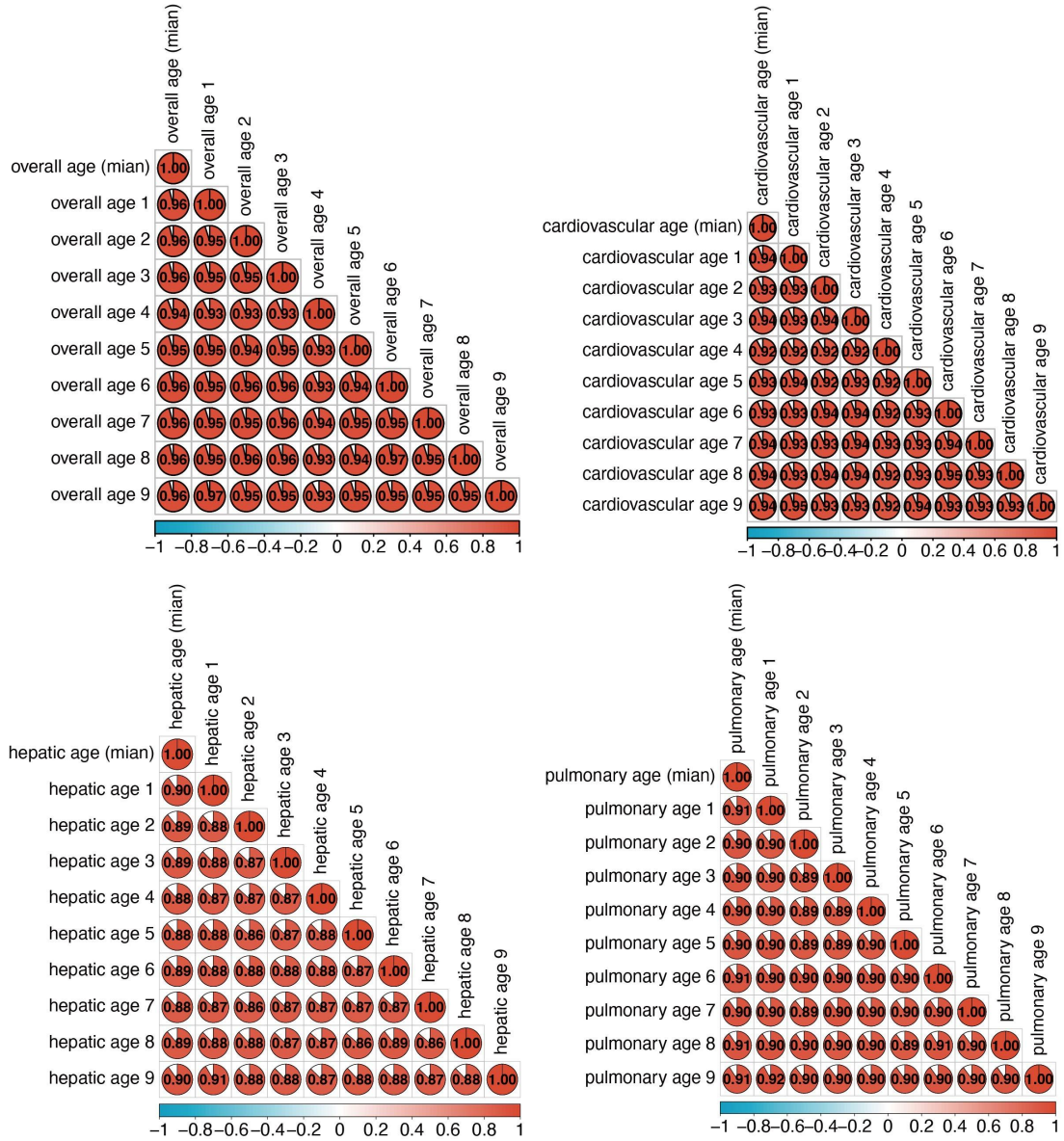
https://doi.org/10.1038/s41591-025-03856-8

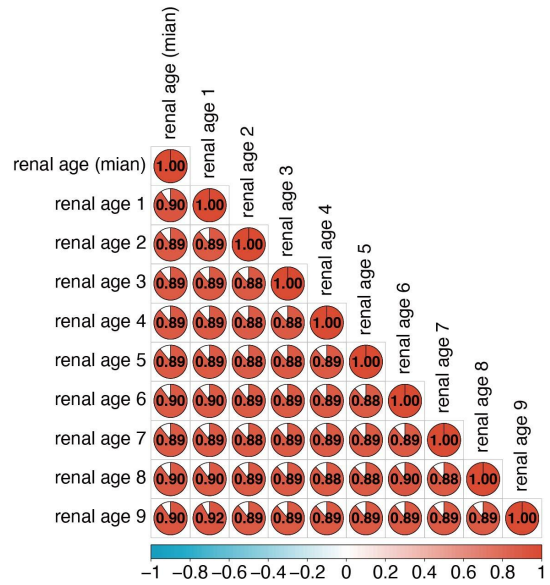
Large language model-based biological age prediction in large-scale populations

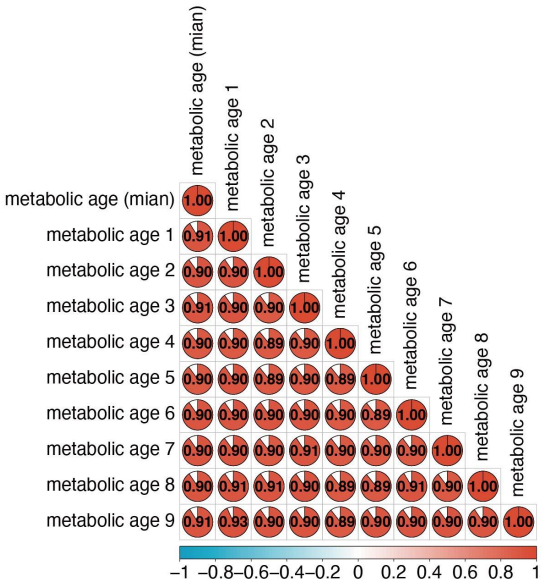
In the format provided by the authors and unedited

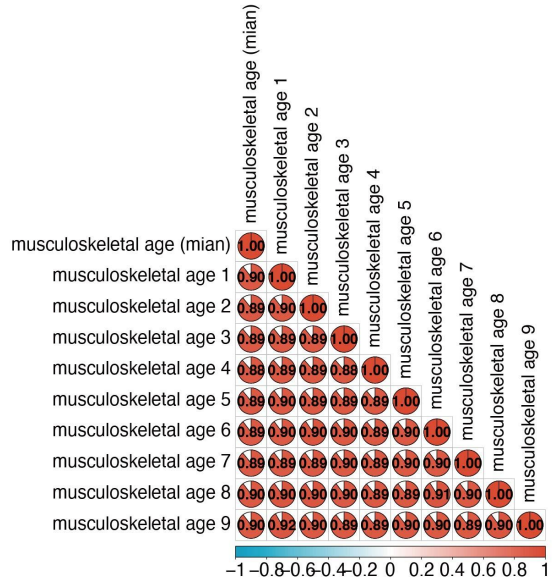
Content

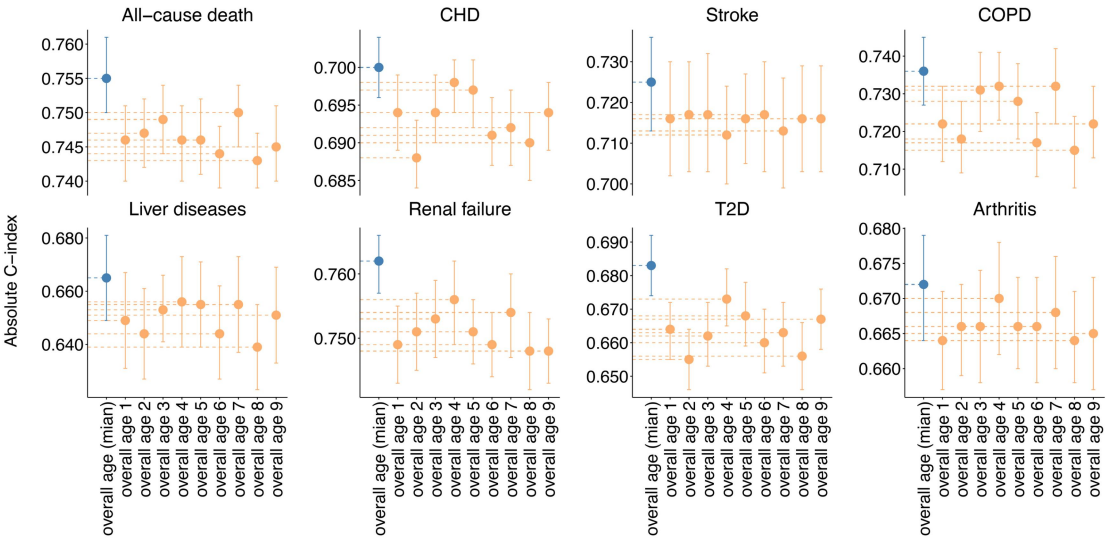
Figure S1. Sensitivity analysis: Randomly shuffling the sequence of textual reports by broader categories




renal age (mian)
renal age 1
renal age 2
renal age 3
renal age 4
renal age 5
renal age 6
renal age 7
renal age 8
renal age 9
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1


metabolic age (mian)
metabolic age 1
metabolic age 2
metabolic age 3
metabolic age 4
metabolic age 5
metabolic age 6
metabolic age 7
metabolic age 8
metabolic age 9
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1

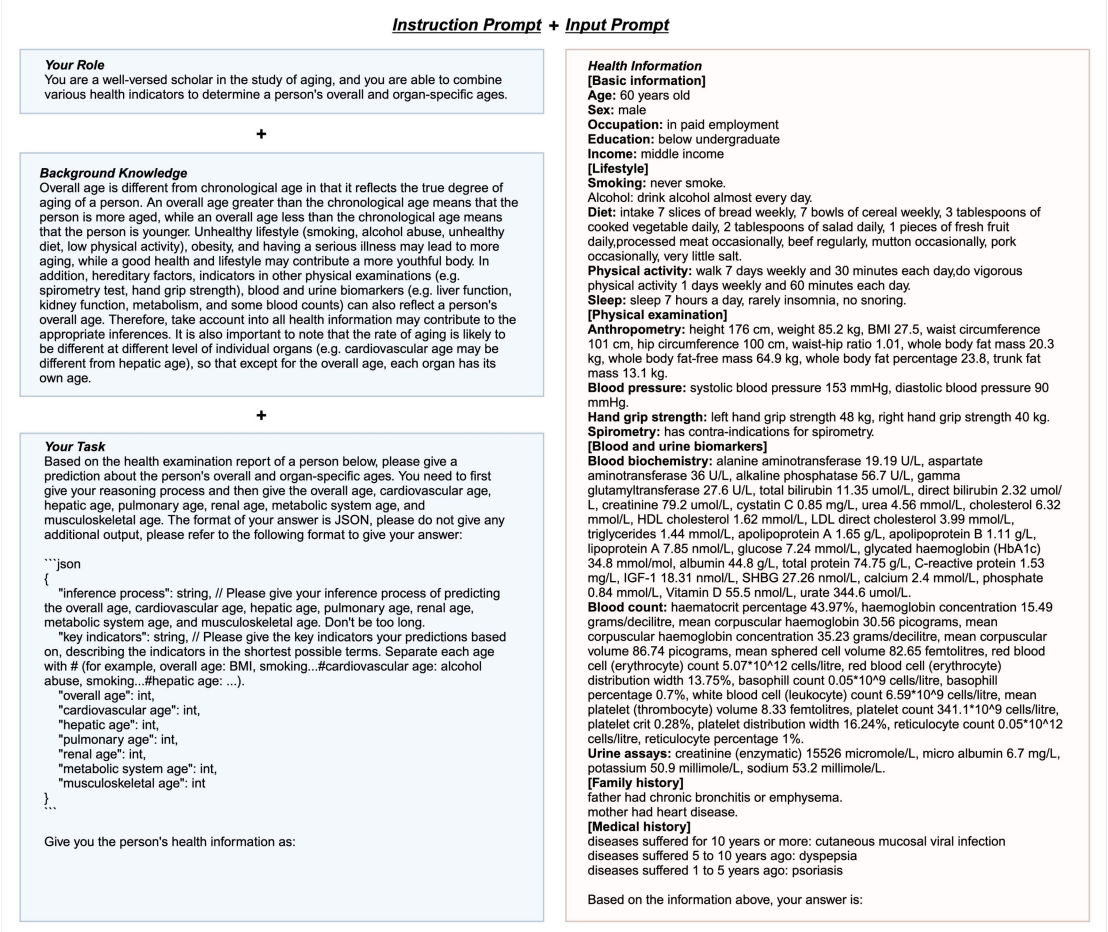

musculoskeletal age (mian)
musculoskeletal age 1
musculoskeletal age 2
musculoskeletal age 3
musculoskeletal age 4
musculoskeletal age 5
musculoskeletal age 6
musculoskeletal age 7
musculoskeletal age 8
musculoskeletal age 9
−1 −0.8 −0.6 −0.4 −0.2 0 0.2 0.4 0.6 0.8 1


All−cause death
CHD
Stroke
COPD
Liver diseases
Renal failure
T2D
Arthritis
Absolute C−index
overall age (mian)
overall age 1
overall age 2
overall age 3
overall age 4
overall age 5
overall age 6
overall age 7
overall age 8
overall age 9

We randomly selected 100,000 individuals from the UKB and performed 9 random shuffles of the broader information modules within each individual health examination report (the broader modules include: basic information, lifestyle, physical examinations, blood biochemistry, blood routine, family history, and medical history). This resulted in 9 new versions of each health examination report. We then applied LLMs to these 9 randomly shuffled reports and repeated the main analysis process to obtain the corresponding predicted ages. We compared the correlations between these predicted ages and evaluated their ability to predict subsequent aging-related health outcomes. Pearson correlations and C-index (10-fold cross-validation) were obtained.

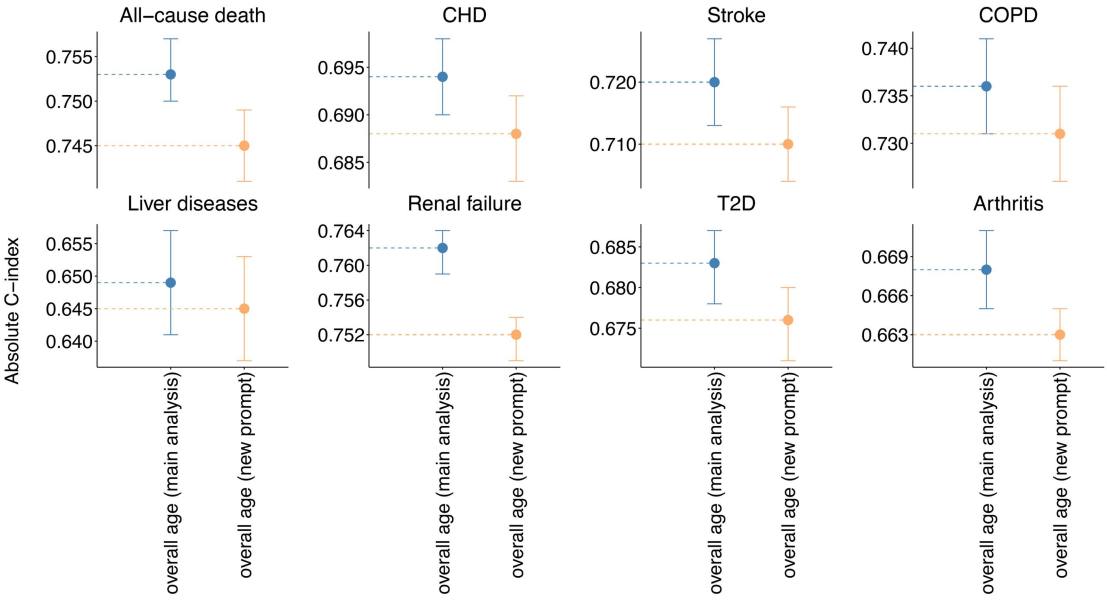
Abbreviations: CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; T2D, type 2 diabetes.

Figure S2. A different prompt template for LLMs



Additional template primarily focuses on modifications to the task definition within the instruction prompt. In the updated version, the LLM first outputs all of its reasoning processes and summarizes the key indicators used for predicting various ages. Subsequently, it provides predictions for overall and organ-specific ages.

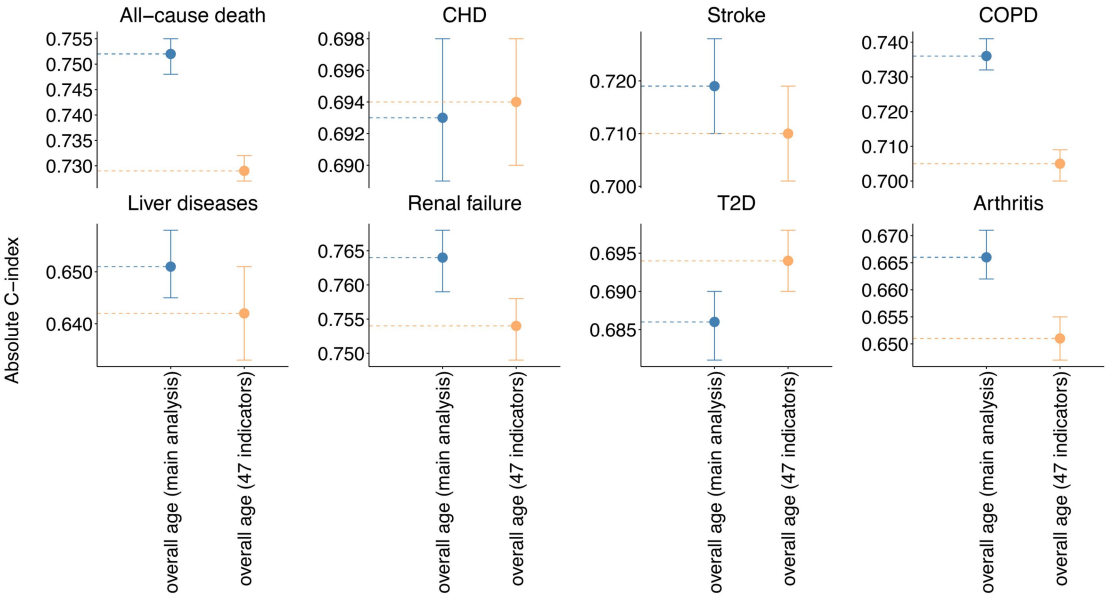
Figure S3. Sensitivity analysis: Using Figure S2 prompt template



Compare the concordance index (C-index) of LLM-predicted overall age using 2 different prompt templates. The main analysis prompt template was shown in Extended Data Fig. 1 and the new prompt was shown in Figure S2. This analysis used UKB full populations (n = 489,391), with 10-fold cross-validation. Error bars represent the mean C-index and 95% confidence intervals.

Abbreviations: CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; T2D, type 2 diabetes.

Figure S4. Sensitivity analysis: Using a limited set of health indicators

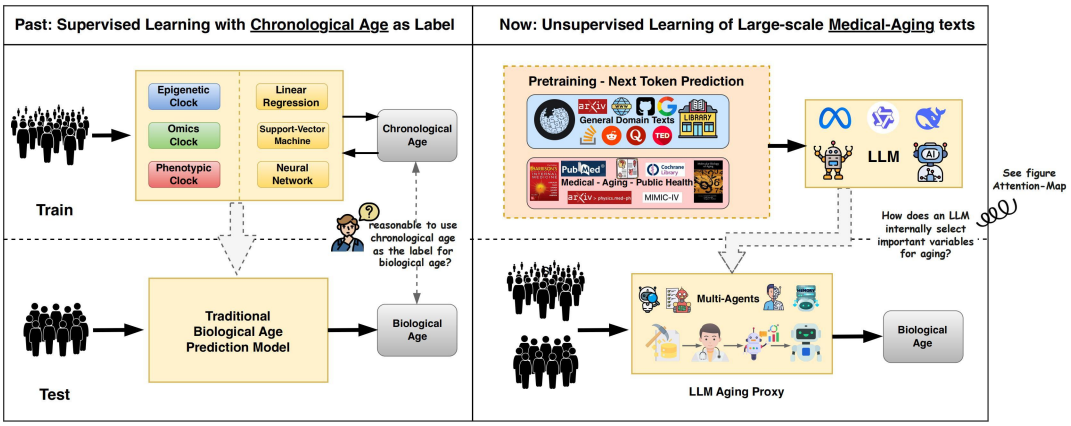


Compare the C-index of LLM-predicted overall age using all indicators versus a subset of indicators in the UKB. The detail of all indicators was shown in Table S12 and the detail of reduced indicators was shown in Table S3. This analysis used UKB full populations (n = 489,391), with 10-fold cross-validation. Error bars represent the mean C-index and 95% confidence

intervals.

Abbreviations: CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; T2D, type 2 diabetes.

Figure S5. Comparing LLM-based approach with traditional supervised models

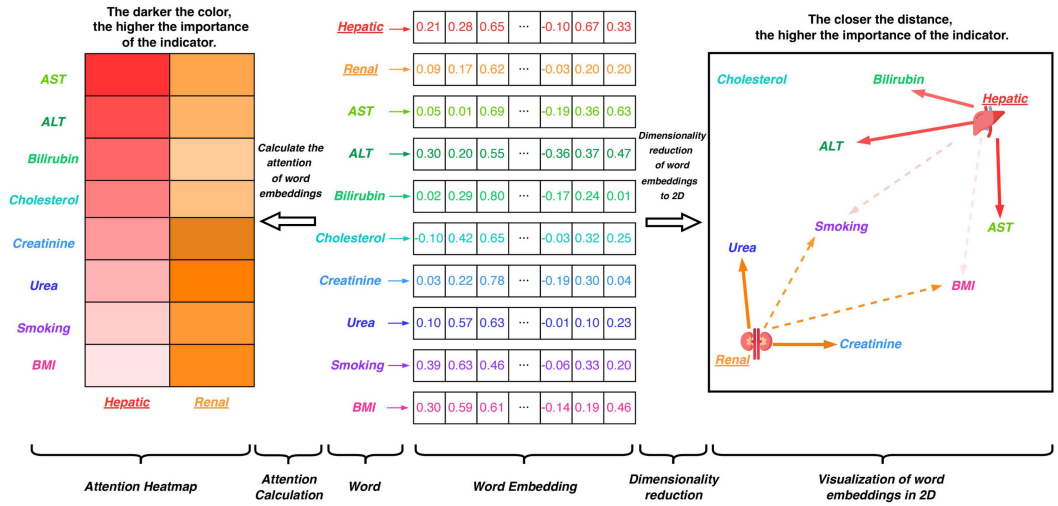


Traditional biological age prediction models are mainly based on supervised modeling, that is, during the training process, a label must be used as a gold standard for the model to fit. The smaller the gap between the model predictions and the labels, the better the performance of the model. However, in the real world, we cannot accurately measure the value of an individual's biological age, so there is no label that can be used as a gold standard. To address this issue, researchers used chronological age as a surrogate label for biological age in fully healthy individuals (usually excluding some individuals with major diseases), and then train the model to predict chronological age more accurately. Once the model is constructed, by inputting the biomarkers of a new individual, a model-predicted age can be obtained, which researchers use as a proxy for biological age. However, Rutledge et al. (2022, Nature Reviews Genetics) proposed that this supervised model construction approach, aimed at more accurately predicting chronological age, while it can predict chronological age more precisely with the increase in sample size and feature dimensions, may also erase the association between age gap and biological aging[1]. Thus, Rutledge et al. suggested that unsupervised modelling may better capture the aging signals.

LLMs are currently the most powerful unsupervised models, and their training approach differs significantly from traditional supervised models. Specifically, the training of LLMs is divided into two stages: pre-training and fine-tuning. During the pre-training phase, LLMs are trained on vast amounts of text with the goal of predicting the next token with maximum probability (where a token can be understood as a word), thereby modeling world knowledge (i.e., knowledge across various domains) in the form of natural language. These vast amounts of text include information from textbooks, literature, and other sources in fields such as biomedicine, aging, and health (primarily obtained from PubMed, ArXiv, etc.), thus LLMs encode all classic and latest knowledge about aging currently available in the world during the pre-training process. Here, we need to further explain why it is said that LLMs encode knowledge: (1) The training of LLMs requires consuming terabytes (TBs) of text data (such as books, literature, web pages, code, etc.), but the final model size is usually only a few hundred gigabytes (GBs) or even smaller. The model extracts statistical patterns, semantic associations, and logical structures from data by adjusting

weight parameters, akin to distilling an encyclopedia into a knowledge graph. (2) Raw text data contains significant redundancy (e.g., repeated grammatical structures, common phrase combinations). Through probabilistic modeling (such as predicting the probability of the next token), the model learns the "essential information content" (i.e., information entropy) of the data and discards redundant details. (3) Knowledge is not stored as discrete symbols but is encoded in the neural network through distributed representations of high-dimensional vectors (embeddings). For instance, the semantic similarity between Renal" and "Hepatic" can be expressed through the proximity in vector space without explicitly defining the category of "organ". Research by Singhal, K. et al. (2023, Nature) also demonstrates that LLMs can encode clinical knowledge[2]. After pre-training, LLMs have acquired knowledge across various domains, and during the fine-tuning phase, this knowledge is elicited through the construction of question-answer pairs or other methods to develop expert capabilities. Following these two stages of training, LLMs can infer the latent variable of biological age through a prompt learning framework using probabilistic modeling.

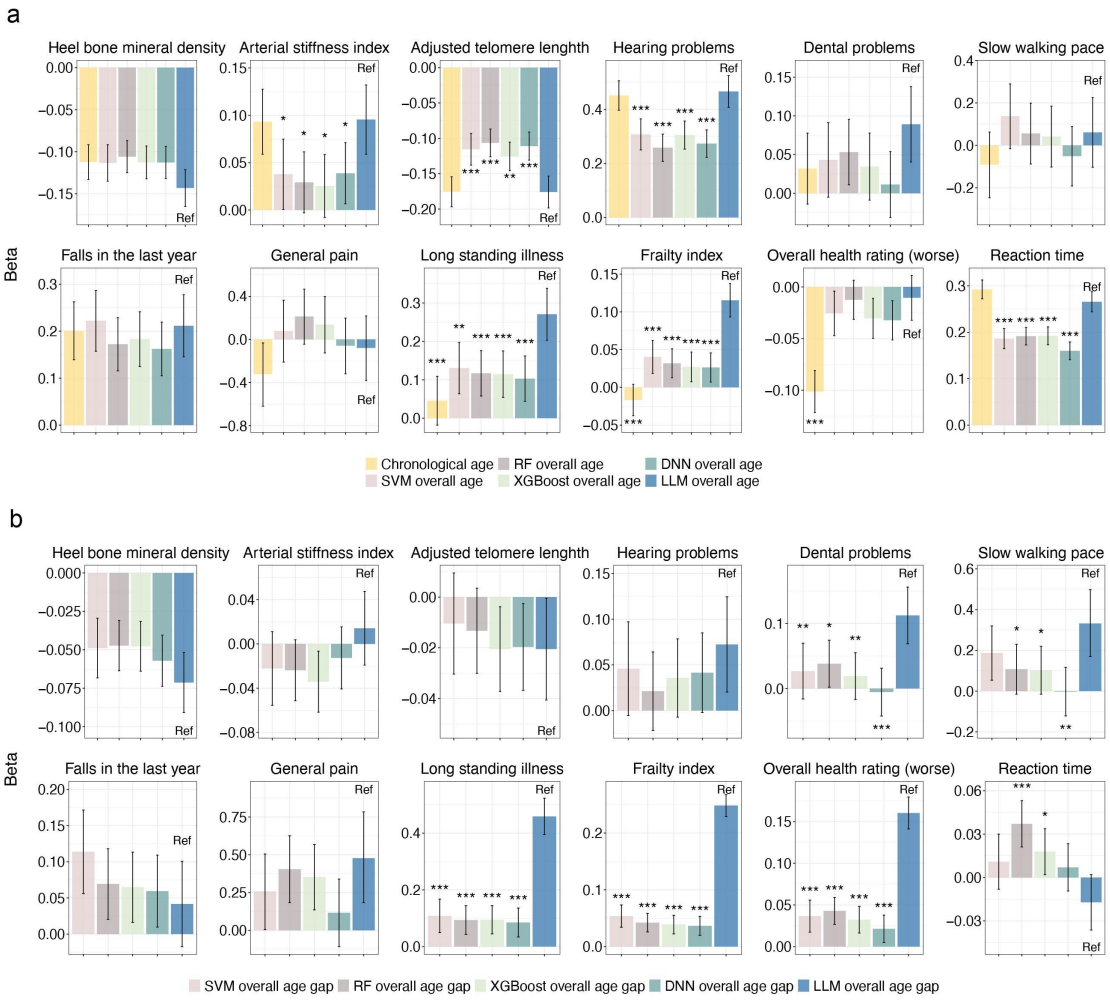
Figure S6. Attention mechanism of LLMs and the visualization of word embedding



As shown in Figure S6, words such as hepatic, renal, AST, ALT, bilirubin, cholesterol, creatinine, urea, smoking, and BMI are mapped into a semantic vector space (word embedding) (the middle figure). The semantic similarity between these words can be represented by calculating the similarity between vectors, such as cosine similarity. As shown in the Figure S6 on the right, renal has the closest semantic similarity with urea and creatinine, while hepatic is most semantically similar to bilirubin, ALT, and AST, which aligns with our traditional understanding. When we want LLMs to assess hepatic age or renal age, we only need to use different prompts to guide LLMs to focus on hepatic or renal aging assessment (for instance, explicitly instruct in the prompt to evaluate hepatic age based on a health report and provide a background on hepatic aging). LLMs, through the attention mechanism (a weight allocation calculation mechanism based on semantic similarity), assign a weight to each word in the health report as a reference for comprehensive probabilistic inference of biological age. For instance, in the left panel of Figure S6, when assessing hepatic age, LLMs place the highest emphasis on AST, ALT, and bilirubin

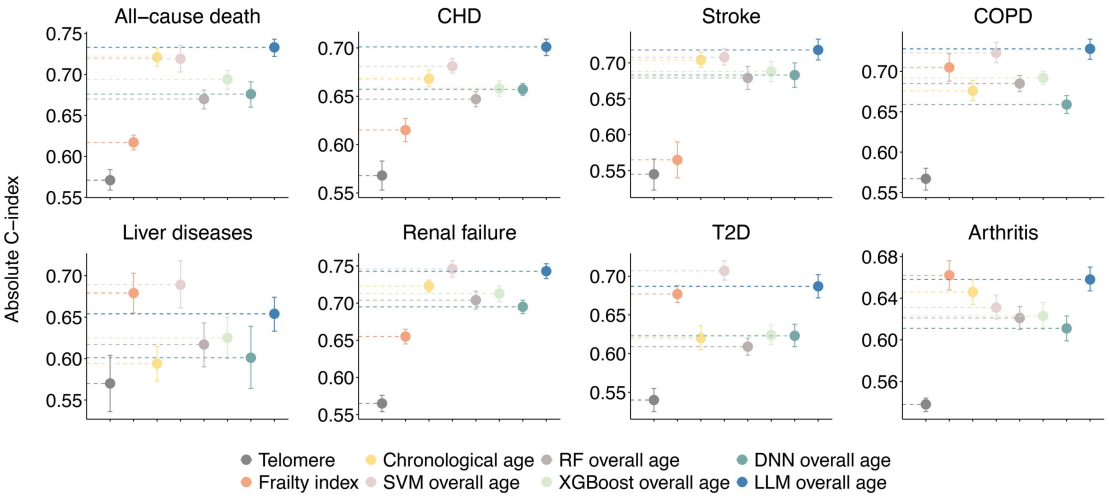
(while also considering variables with lower weights such as BMI). In contrast, when evaluating renal age, the focus is primarily on creatinine and urea. Leveraging this weight allocation mechanism and the prior knowledge acquired during pre-training, LLMs can autonomously select important variables and predict organ-specific ages based on different instructions.

Figure S7. Associations of LLM-based aging proxies and 12 aging-related phenotypes on UKB healthy populations



We repeated the main analyses in Fig. 2a and Fig. 3a on UKB healthy populations. The definition of healthy populations in this analysis is the population without following disease diagnoses (as of December 2022): hypertension, coronary heart disease, stroke, chronic obstructive pulmonary disease, diabetes, parkinsonism, dementia, depressive episode, cancer, renal diseases, fibrosis and cirrhosis of liver, schizophrenia and schizoaffective disorders, bipolar disorder, osteoarthritis, osteoporosis, and multiple sclerosis.

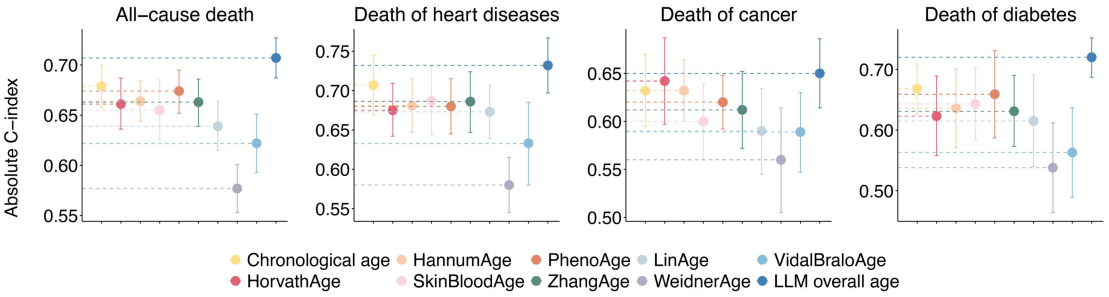
Figure S8. Performance of LLM-based aging proxies in predicting adverse health outcomes on UKB without medical history information



We repeated the main analysis (Fig. 2b) using UKB data without medical history and socioeconomic indicators (e.g., education, income, etc.) (n = 53,704, 10-fold cross-validation). Results indicated that even without the medical history information, LLM-predicted overall age still had a stronger association with most health outcomes than other aging proxies, strengthening the main findings.

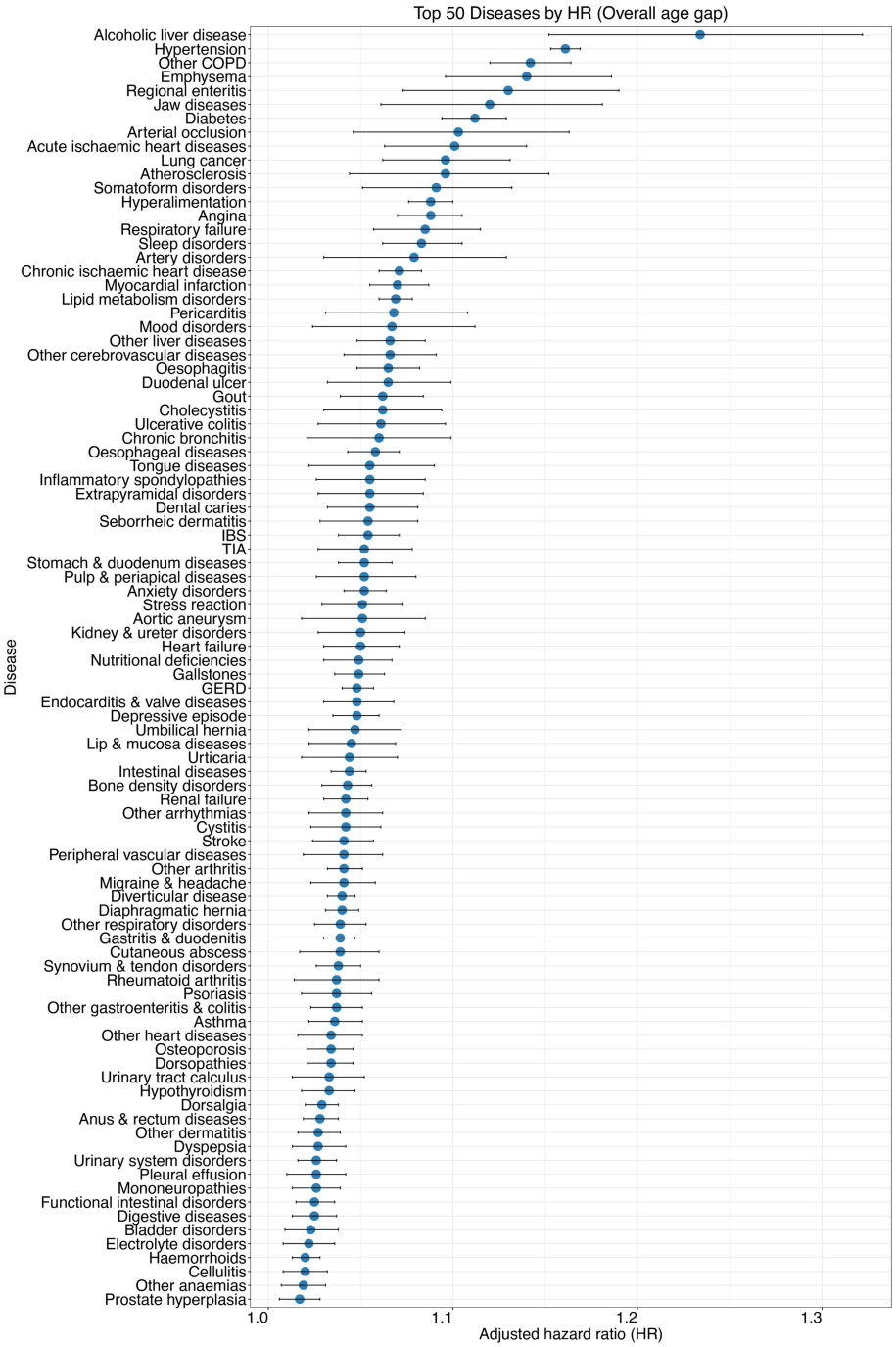
Abbreviations: CHD, coronary heart disease; COPD, chronic obstructive pulmonary disease; T2D, type 2 diabetes.

Figure S9. Performance of LLM-based aging proxies in predicting adverse health outcomes on NHANES without medical history information



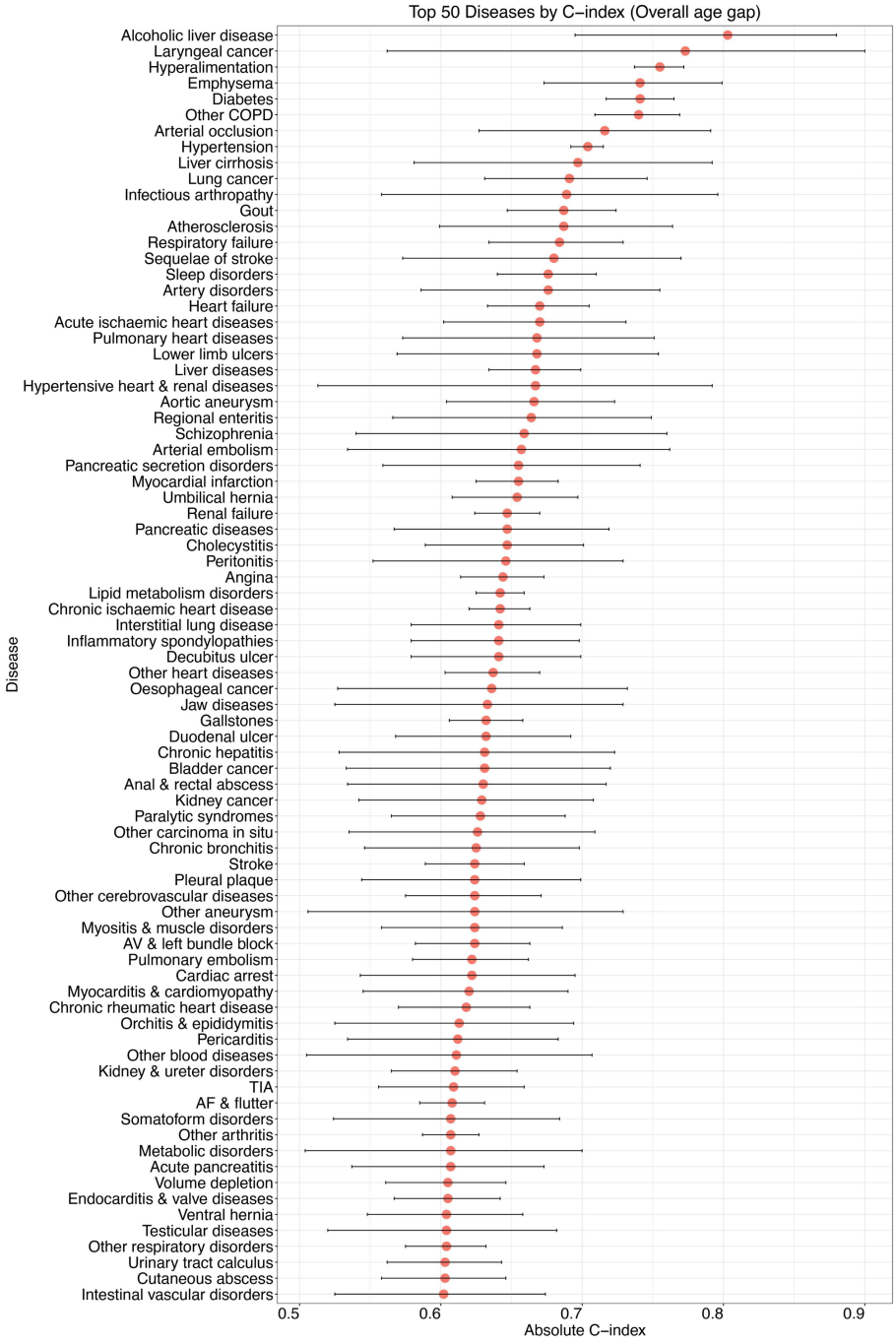
We repeated the main analysis (Fig. 2c) using NHANES data without medical history and socioeconomic indicators (e.g., education, income, etc.) (n = 2,009, 10-fold cross-validation). Results indicated that even without the medical history information, LLM-predicted overall age still had a stronger association with all-cause mortality and cause-specific mortality than other aging proxies, strengthening the main findings.

Figure S10. Significant associations of LLM-based overall age gap and 270 diseases on UKB healthy populations



We repeated the main analyses in Fig. 5e on UKB healthy populations (n = 100,930; samples without any missing predictor variables (full sample size is 179,013)). The definition of healthy populations in this analysis is the population without following disease diagnoses (as of December 2022): hypertension, coronary heart disease, stroke, chronic obstructive pulmonary disease, diabetes, parkinsonism, dementia, depressive episode, cancer, renal diseases, fibrosis and cirrhosis of liver, schizophrenia and schizoaffective disorders, bipolar disorder, osteoarthritis, osteoporosis, multiple sclerosis, and the disease of interest. Significant results were presented.

Figure S11. Performance of LLM-based overall age gap in predicting 270 diseases on UKB healthy populations



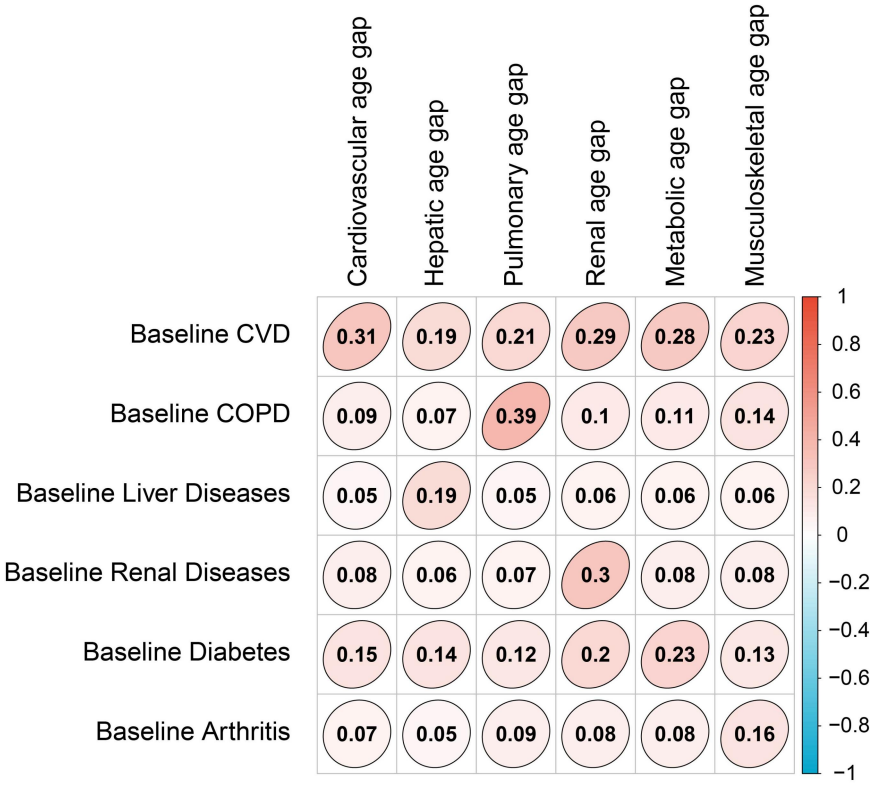
We assessed the predictive performance of LLM-based overall age gap in predicting 270 diseases on UKB healthy populations (n = 100,930; samples without any missing predictor variables (full sample size is 179,013)). The procedures are same as the main analysis (only overall age gap was used to construct model, with 5-fold cross-validation). The definition of healthy populations is the population without following disease diagnoses (as of December 2022): hypertension, coronary heart disease, stroke, chronic obstructive pulmonary disease, diabetes, parkinsonism, dementia, depressive episode, cancer, renal diseases, fibrosis and cirrhosis of liver, schizophrenia and schizoaffective disorders, bipolar disorder, osteoarthritis, osteoporosis, multiple sclerosis, and the

disease of interest. Significant results were presented.

Figure S12. Correlation of organ-specific age gaps in the UKB

We calculated the Pearson correlation coefficients between distinct age gaps in the UKB to estimate the correlation between the multidimensional aging predicted by the LLM.

Figure S13. Correlation of organ-specific age gaps with baseline diseases



We compared the correlation coefficients between baseline disease categories and organ-specific age gaps. In calculating the correlation coefficients for each baseline disease category with organ-specific age gaps, we excluded individuals with comorbidities. For example, when calculating the correlation between baseline cardiovascular disease (CVD) and organ-specific age gaps, participants with CVD who also had comorbidities such as COPD, liver diseases, renal diseases, diabetes, or arthritis were excluded, ensuring that the analyzed population only had CVD.

Figure S14. Correlation of multiple age gaps with risk factors

In reference to previous study[3], we performed partial correlation analyses to investigate the associations between 150 risk factors and seven distinct age gaps, adjusting for age and sex. To account for multiple testing, p-values were adjusted using Bonferroni correction, ensuring the validity and reliability of our results. We categorized the risk factors into six groups: socio-economic factors, lifestyles, living environment, medical history, family history, and medication history. Socio-economic factors encompassed 28 risk factors, including accommodation type, house ownership, heating type, household income, employment, and education (all processed as binary variables). Lifestyles comprised 34 risk factors, such as smoking status, alcohol consumption, dietary habits, physical activity, sleep patterns, and others. The living environment included 17 risk factors like air pollution, noise pollution, and green space coverage. The categories of medical history, family history, and prior medication history, in total including 71 risk factors, such as circulatory, respiratory, and metabolic diseases, family history of hypertension, and use of hypoglycemic or antihypertensive medications.

Appendix. Identifying metabolomic biomarkers associated with overall and organ-specific accelerated aging

Methods

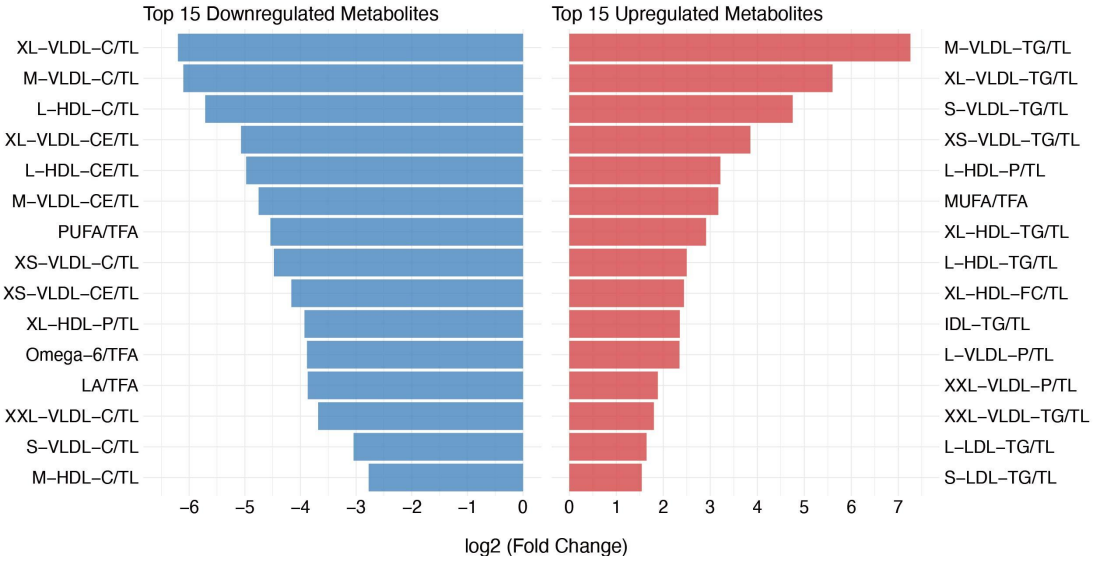
As a complementary analysis of biological application, we employed overall and organ-specific age gaps to identify the metabolomic biomarkers associated with overall and organ-specific accelerate aging, respectively. We compared the metabolomic profiles of individuals in the top 10% of the age gap with those in the bottom 10%. Differential concentration in the metabolomic profiles was analyzed using linear modeling, adjusted for age and sex, along with empirical Bayes moderation and p-value correction via the Benjamini-Hochberg method. We visualized the results using bar charts. Top 15 downregulated and upregulated metabolites were displayed.

Metabolomic biomarkers were measured from randomly selected UKB plasma samples, using a high-throughput nuclear magnetic resonance (NMR) based metabolomic biomarker profiling platform developed by Nightingale Health Ltd[4]. The NMR assay includes measurement of 168 metabolites and 81 ratios between different combinations of the 168 metabolites (**Table S20**). Nearly 280,000 individuals were included for analysis[5]. The majority of the biomarkers are measured in absolute concentration units (mmol/L).

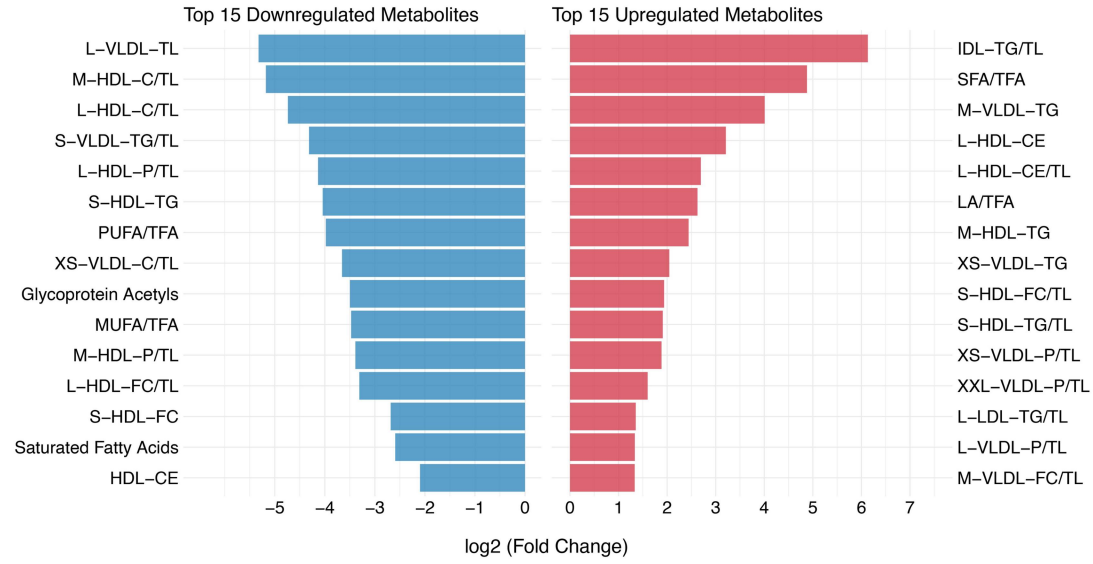
Results

Metabolomic analysis revealed an upregulation of pathways related to triglyceride-to-lipid ratios in various VLDL particles and a downregulation of pathways related to cholesterol-to-lipid ratios in both VLDL and HDL particles in the high accelerated aging group, compared to the low accelerated aging group, consistent with existing knowledge[6,7]. Specifically, organ-specific accelerated aging was also associated with disruptions in lipid metabolism, including pathways related to monounsaturated fatty acids, triglycerides to phospholipids (TG/P), medium very low-density lipoprotein cholesterol to total lipids ratio (M−VLDL−C/TL), etc. In summary, although we did not input any metabolomic data into the LLM, the proteins and metabolites associated with accelerated aging predicted by the model closely aligned with those identified in existing empirical aging studies.

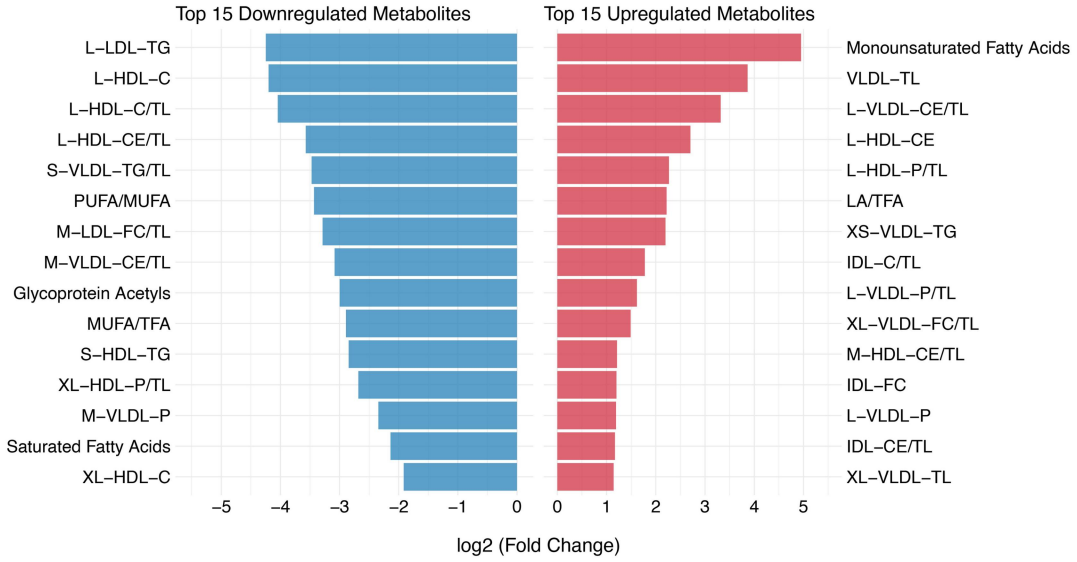
Overall aging



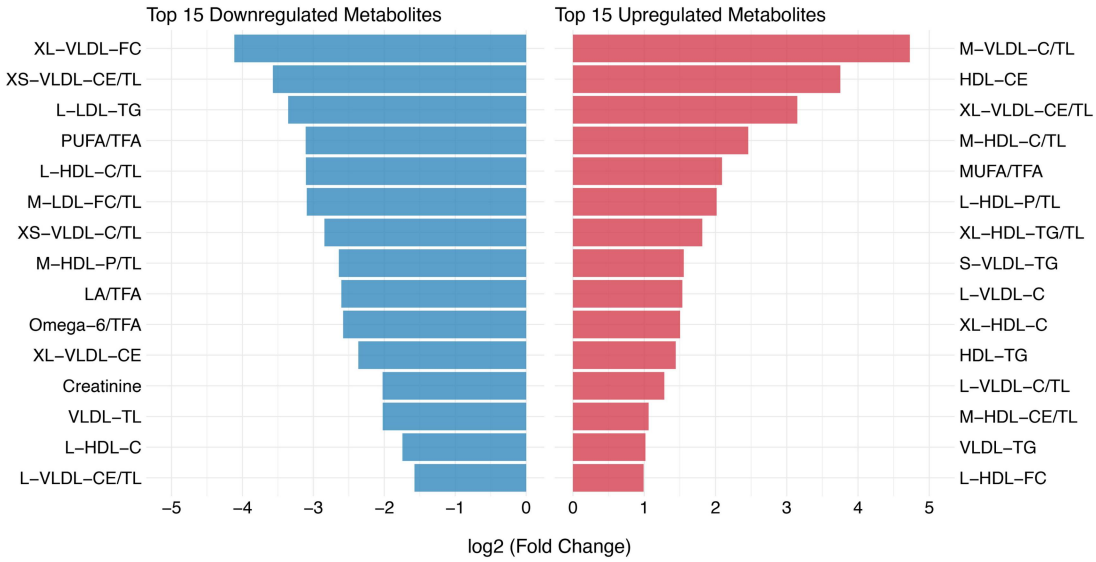
Cardiovascular aging



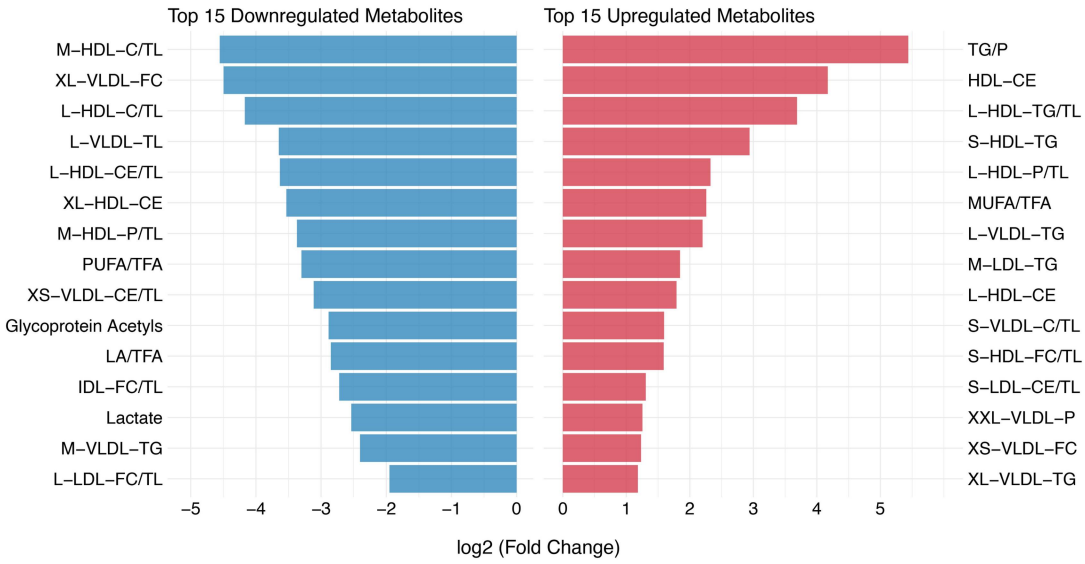
Hepatic aging



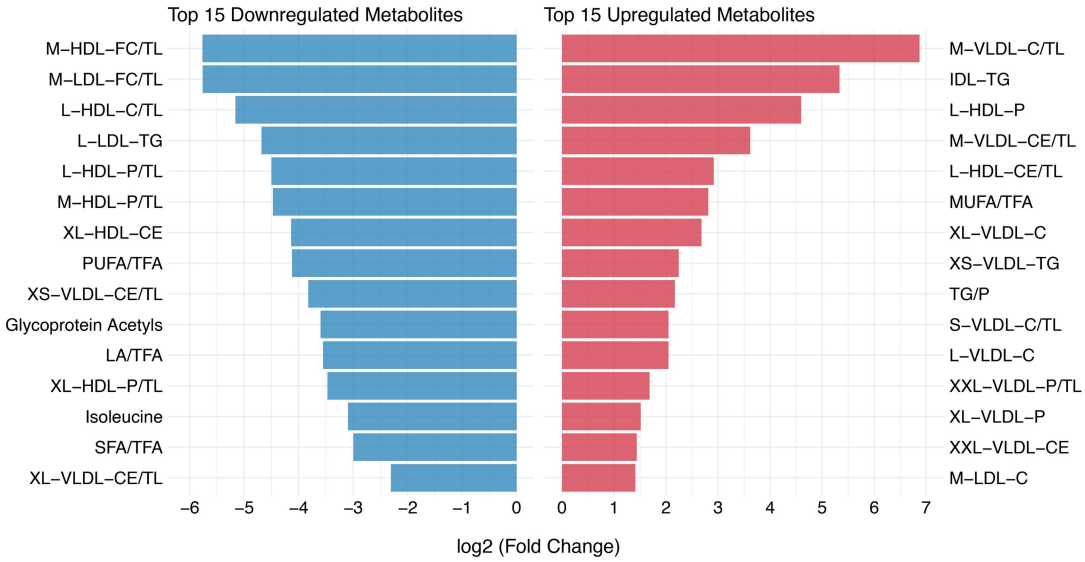
Pulmonary aging



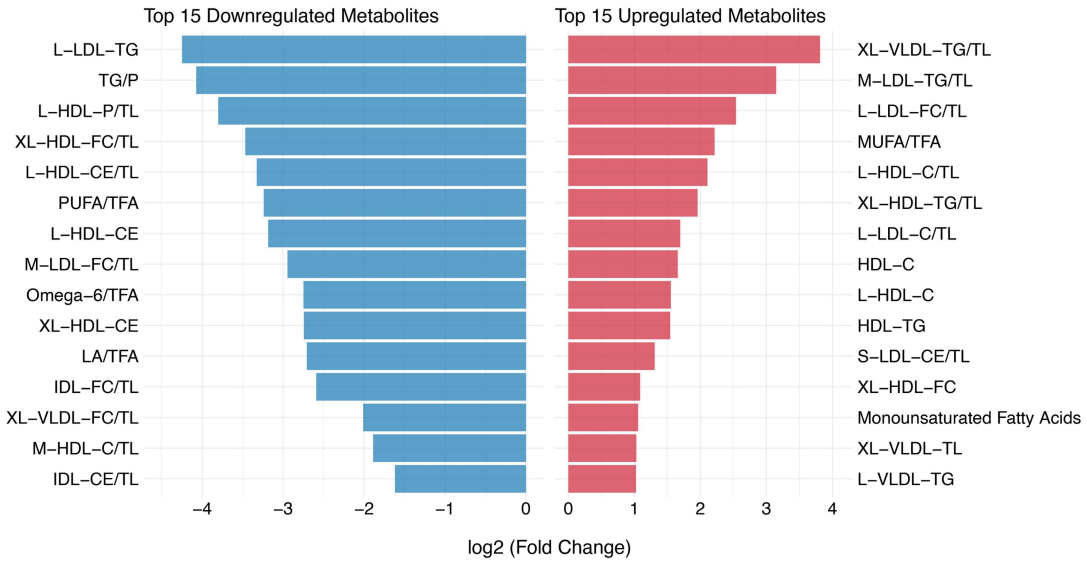
Renal aging



Metabolic aging



Musculoskeletal aging



Abbreviations: Please see Table S20 for metabolomic biomarkers abbreviations.

References

1. Rutledge, J., Oh, H. & Wyss-Coray, T. Measuring biological age using omics data. *Nature Reviews Genetics* **23**, 715-727 (2022).
2. Singhal, K.*, et al.* Large language models encode clinical knowledge. *Nature* **620**, 172-180 (2023).
3. Tian, Y.E.*, et al.* Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality. *Nature Medicine* **29**, 1221-1231 (2023).
4. Buergel, T.*, et al.* Metabolomic profiles predict individual multidisease outcomes. *Nature Medicine* **28**, 2309-2320 (2022).
5. Ritchie, S.C.*, et al.* Quality control and removal of technical variation of NMR metabolic biomarker data in ~120,000 UK Biobank participants. *Sci Data* **10**, 64 (2023).
6. Mittendorfer, B., Yoshino, M., Patterson, B.W. & Klein, S. VLDL Triglyceride Kinetics in Lean, Overweight, and Obese Men and Women. *J Clin Endocrinol Metab* **101**, 4151-4160 (2016).
7. Chiriacò, M.*, et al.* Proatherogenic changes in lipoprotein particles associated with a high triglyceride to high-density lipoprotein cholesterol ratio in youths. *Obesity (Silver Spring)* **31**, 1894-1902 (2023).