LongevityBench: Are SotA LLMs ready for aging research?

Alex Zhavoronkov[1,2], Denis Sidorenko[1], Vladimir Naumov[1], Stefan Pushkov[1], Diana Zagirova[1], Vladimir Aladinskiy[1], Alex Aliper[1], Fedor Galkin[1,†]

[1] Insilico Medicine, Abu Dhabi, UAE; [2] Buck Institute for Research on Aging, Novato, California
† Corresponding author: f.galkin@insilico.com

Abstract

Aging is a core biological process observed in most species and tissues, which is studied with a vast array of technologies. We argue that the abilities of AI systems to emulate aging and to accurately interpret biodata in its context are the key criteria to judge an LLM's utility in biomedical research. Here, we present LongevityBench — a collection of tasks designed to assess whether LLMs grasp the fundamental principles of aging biology and can use low-level biodata to arrive at phenotype-level conclusions. The benchmark covers a variety of prediction targets including human time-to-death, mutations' effect on lifespan, and age-dependent expression patterns. It spans all common biodata types used in longevity research: transcriptomes, DNA methylation profiles, proteomes, genomes, clinical blood tests and biometrics, as well as natural language annotations. After ranking state-of-the-art LLMs using LongevityBench, we highlight their weaknesses and outline procedures to maximize their utility in aging research.

1. Introduction

Early applications of deep learning methodologies for the study of aging were aimed at discovering its biomarkers and measuring its intensity within the aging clock paradigm (Putin et al., 2016). The 2010s saw a boom of deep aging clocks used to identify previously overlooked aging patterns in such diverse biodata types as DNA methylation (DNAm) (Fedor Galkin et al., 2020), gut flora composition (Galkin et al., 2020), mRNA expression (Mamoshina et al., 2018b), facial photography (Bobrov et al., 2018), clinical blood tests (Mamoshina et al., 2018a), and psychological evaluations (Zhavoronkov et al., 2020). These models were used for controlling batch effects in aging studies (Mamoshina et al., 2019), for target and mechanism discovery (Zhavoronkov and Mamoshina, 2019), and for anti-aging clinical trial organization (Zhavoronkov

and Mamoshina, 2019). Later, more complicated methodologies were developed to enable generative, rather than descriptive, capabilities of aging deep learning models. Such models could be used to learn joint representations of aging processes from disparate biodata types to extrapolate them and identify compounds that could elicit the desired pro-longevity transcriptomic changes (Aliper et al., 2022; Shayakhmetov et al., 2020). Most recently, the transformer LLM architectures have been adopted to manipulate low-level biodata and enable multi-modal omics comprehension, conditional generation, and simulation of aging processes in multiple species, as presented in the Precious line of models (Galkin et al., 2024; Sidorenko et al., 2024; Urban et al., 2023).

The vector of dedicated biodata-focused AI development continues to this day and has become intertwined with the progress in the general-purpose AI systems. Commercially released LLMs, such as ChatGPT, Claude, Grok, Gemini have deeply affected the research community. Today, biomedical researchers use AI for a variety of professional tasks, which has resulted in a major increase in their productivity as academic authors (Kusumegi et al., 2025). AI applications help scientists with routine tasks, write texts, provide references, write code, and ultimately save the researchers' time which can be allocated to planning new experiments, exploring findings, and hypothesis generation. Some envision an even deeper integration of AI into the scientific process and propose systems that can handle end-to-end research projects by assigning each stage to a specialized intelligent agent (Schmidgall et al., 2025). Such systems are demonstrated to navigate quite complex tasks and provide realistic-looking publication drafts.

However, the grasp of basic biological ground truth by these LLMs and agentic systems remains a subject of much debate. The skepticism toward AI's involvement in biomedical research is most commonly expressed on professional forums and in private conversations, with peer-reviewed critical pieces rarely offering solutions to mitigate the negative impact AI may have on scientific progress (Hicks et al., 2024; Messeri and Crockett, 2024; Singh, 2025, p. 202).

We argue that the research community's worries can be remedied by a disciplined approach to model evaluation. Benchmarks provide an empirical foundation for calibrating user trust: a model that fails to correctly interpret standard biodata cannot be trusted as a research companion, regardless of how fluently it communicates.

The field of aging biology offers a wide range of tests that can be used to judge an LLM's grasp of the biological ground truth, ability to reason and derive high-order features from low-level information, and thus, evaluate its fitness for takin part in the scientific process. A benchmark set covering varied data types, biological scales, and question formats could be used as a test for the presence of a coherent internal model of life in general, making it not just a way to filter out unreliable LLMs, but also a development target. By steering model training toward mastery in all challenges presented by aging research, we can achieve models operating based on genuine understanding of the biological reality rather than superficial pattern-matching.

To this end, we present LongevityBench, a collection of benchmark datasets designed to gauge the ability of current state-of-the-art (SotA) models to operate in the aging biology domain and extrapolate from molecular measurements to organism-level phenotypes. The benchmark includes tasks pertaining to lifespan, mortality, and age prediction across multiple species and experimental paradigms. Task aimed at evaluating generative output with age as a conditional are particularly challenging and thus may be used as the ultimate test of a model's reasoning capacity and the ability to segregate technical noise from biological signal (Mamoshina et al., 2019). Our library of prompts offers a variety of question formats, allowing us to measure performance in aging research without confounding by query semantics. The inclusion of all major biodata types enables us to determine whether LLMs operate within a unified framework of aging rather than a fragmented collection of domain-specific associations.

We aim to achieve two goals with LongevityBench. First, we seek to create a leaderboard of SotA LLMs to help scientists choose the best base model to serve as their research companion in longevity studies. Second, we put forward principles of responsible research-grade AI development. The presented work is an integral part of our Multi-Modal AI (MMAI) Gym for Science, a proprietary pipeline enabling the training of AI systems capable of tackling advanced problems in drug discovery and development (**Figure 1**). The MMAI environment leverages domain-specific supervised and reinforcement fine-tuning (SFT and RFT) algorithms to ensure a model's alignment with empirically established ground truths and to teach chemical and biological reasoning patterns. As an MMAI Gym module, LongevityBench serves as a rigorous evaluation method for engineers who aim to create LLMs with the capacity to generalize across biological data and provide actionable insights.

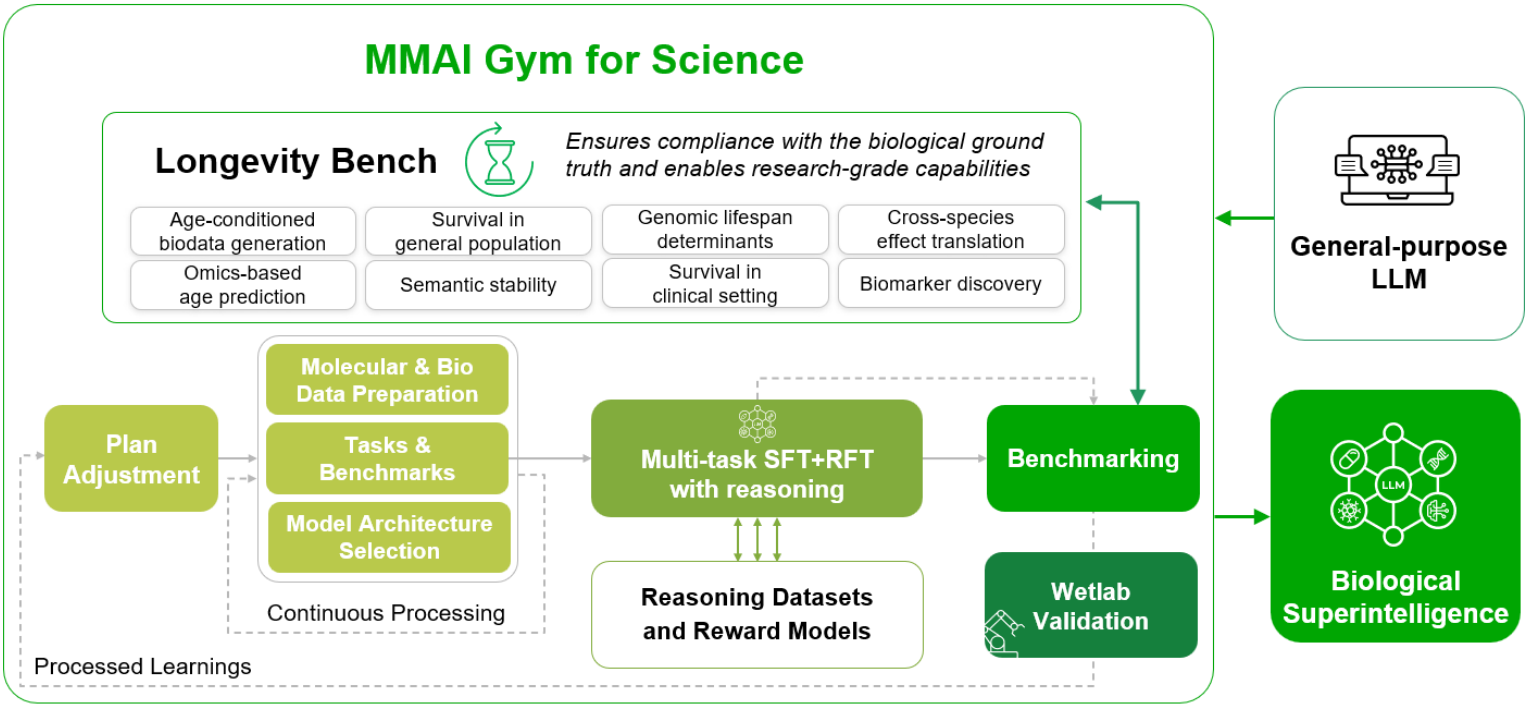

Figure 1. High-level scheme of an LLM training in the Multi-Modal AI (MMAI) Gym for Science. The gym is an iterative pipeline which leverages a variety of deep learning techniques, such as supervised fine-tuning (SFT) and reinforcement fine-tuning (RFT) based on carefully annotated biological and chemical datasets to enhance LLMs' alignment with the empirical ground truth. In this pipeline, LongevityBench serves a crucial role by grounding the training process with a set of tests aimed at evaluating a model's understanding of aging biology.

2. Methods

2.1. Benchmark Datasets

To create LongevityBench, we used multiple publicly available data sources to compose structured prompts in formats that may be processed by LLMs and enable straightforward metrics calculations. The datasets were selected to contain relevant aging-related annotations, which include chronological age, group lifespan, and survival information. Each data source was prepared in several alternative ways that differ by either the depth of the context provided to the model in the prompt or the format in which the model is instructed to provide the answer. The presented collection features prompts that request the model to carry out the regression task (e.g. specify the exact time-to-death), binary classification (e.g. will an individual survive >10 years), pairwise selection (e.g. which individual will live longer) with binary or ternary options, depending on whether tied pairs are included in the collection, and multiple choice (e.g. will the individual survive <5, 5-10, >10 years). For all multiple-choice formats, the order of presented options was randomized to prevent biased learning at supervised fine-tuning (SFT) stages. All prompts were formatted as XML-like structures with tags separating different context sections, enabling easy control over the amount of context an LLM received during testing. All prompts were supplemented with a "`<follow_up>`" section to enable manual curation and sanity checks, which was removed during SFT and benchmark stages. See **"Supplementary materials"** for prompt samples.

See below the details of data collection, partition, and processing for each category with sample prompts for each supported format.

General population survival

We used the NHANES-IV dataset for waves collected in 1999-2017. The two oldest waves (1999, 2001) were separated into a test set, with the rest forming the training set.

For the binary format, we asked LLMs if an individual survived for at least 10 years post-examination. The training set contained 10,000 individuals with equal sex and age group proportions, among them 3000 individuals were confirmed deceased by the end of observation with the 127±54 (mean±std) month survival time. The test set contained 8976 individuals balanced for sex and age group, among them 1356 confirmed deceased with 194±61 survival time. The natural text annotations for the prompts were prepared using 54 variables from the harmonized NHANES dataset (CC BY 4.0 licence) (Nguyen et al., 2023) describing demographic parameters of each individual, details of their health history and physical fitness. The blood biomarker data to supplement the prompts was obtained through the `pynhanes` package (MIT licence) (Tim Pyrkov, 2022). The mean number of blood markers per sample was 28, in the training set and 29 in the test set.

Apart from the binary formats, we prepared pairwise time-to-death (TTD) comparisons with 20,000 pairs of NHANES participants featuring 8606 unique individuals in the training set, and 15,542 pairs with 4697

unique individuals in the test set. Two types of prompts were created for singular TTD prediction: (i) binned with options specified as “0-5 years”, “5-10 years”, “10-15 years”, and "15+years”, or (ii) regression format, in which an LLM was required to predict the exact TTD of a participant in months.

Aging trajectories

We used the Open Genes (Rafikova et al., 2024) database (MPL-2.0 licence) to compose prompts that asked an LLM to assess the directionality of a gene’s expression change in human tissues. Based on this dataset, we prepared 2,952 prompts, of which 647 were split into a test set based on the protein family assignment. Among the total set, the majority of prompts represented findings in the human muscle (N=964), CD8+ T-cells (N=718), and frontal cortex (N=350). The prompts represent mostly mRNA-level expression changes (N=2681) and have a balanced proportion of genes with an upward or downward aging trajectory (54% vs 46%). Among the 1953 unique genes found in the Open Genes database, 157 were annotated with findings in model organisms, such as mouse, rats, and others. The prompts contain additional sections providing context about a masked gene’s function, based on Gene Ontology (GO), and human tissue localization, based on Protein Atlas.

Multi-mutant lifespan

We used the Synergy Age (Bunu et al., 2020) (CC BY 4.0) database to prepare prompts featuring genetic interventions into model organisms with known effect on the lifespan. The 1509 database entries featuring *Caenorhabditis elegans* experiments with 351 unique genes formed the training set, while the 48 *Drosophila melanogaster* and 16 *Mus musculus* experiments formed the test set containing 43 unique genes. The pairwise prompts required an LLM to assess whether a single-mutant or a double-mutant organism had a longer lifespan. In the training set, the double mutants were the answer 50% of the time, while in the test set, they constituted only 41% of correct answers. For the regression format, the model was asked to provide the exact percentage increase in lifespan for a double-mutant organism compared to the wild type (WT). In the training set, 68% of all such prompts featured positive effects on lifespan but in the test set only 39% of double mutants had increased life duration. As part of our ablation study, the pairwise prompts were prepared in the minimal format containing only lifespan durations of single mutants as extra context, or in the extended format featuring additional GO annotations and experimental context pulled from source publications.

Cancer survival

We used TCGA Open Access data, as defined in the NIH GDS and NCI GDS policies, to prepare prompts featuring RNAseq findings, patient diagnosis and treatment information, to predict patients’ progression-free survival interval (PFS). The split was prepared using different tumor types with BLCA, COADRED, ESCA, GBM, HNSC, LIHC, PRAD, STAD, and THCA forming the training set, while KICH, LUAD, and SKCM cohorts formed the test set.

All prompts were formatted as pairwise choice with randomly selected two patients with the same tumor type. The training set was composed of 8,777 such pairs and the test set — of 1,633 pairs of patients. Each prompt is supplemented with processed RNAseq data. For each patient we ran single-sample gene set enrichment analysis (ssGSEA) against the Reactome pathway database to select the most characteristic pathways for each patient, based on normalized enrichment scores (NES) statistics. During the ssGSEA stage, we used fold changes between a patient's normal and malignant tissues to create preranked lists. The prompts contained up to 5 pathways with the most dissimilar NES between patients in a pair, and up to 5 pathways with extreme NES scores across all patients ($<5^{th}$ or $>95^{th}$ percentiles of the total distribution). Additionally, the prompts displayed up to 10 most highly expressed genes for the featured pathways with a positive NES, or 10 genes with the lowest expression values from the pathways with a negative NES. By avoiding exact expression values and focusing on pathway-level annotations, we thus managed to present informative and compact RNAseq context for tumor environments. In the end, the training set featured 603 unique Reactome pathways and 5,417 unique genes, while the test set contained 424 pathways and 3,740 genes.

Methylation aging

We used a collection of Illumina Human Methylation studies released on Gene Expression Omnibus (GEO) described in (Fedor Galkin et al., 2020) to create prompts aimed at measuring the ability of LLMs to process epigenetic signals with minimal processing into chronological age predictions. The training set was composed of blood-derived DNAm records available in GEO submissions: GSE106648, GSE125105, GSE128235, GSE19711, GSE27044, GSE30870, GSE40279, GSE41037, GSE52588, GSE53740, GSE58119, GSE67530, GSE77445, GSE77696, GSE81961, GSE84624, GSE97362. The test set contained samples from: GSE102177, GSE103911, GSE105123, GSE107459, GSE107737, GSE112696, GSE20067, GSE34639, GSE37008, GSE59065, GSE61496, GSE79329, GSE87582, GSE87640, GSE98876, GSE99624. When available, prompts featured information about samples' sex, associated diseases, smoking, BMI, pregnancy, and alcohol use. The prompts were prepared in three formats. For the multiple-choice format, LLMs were asked to choose the by-decade age group of a single sample, ranging from "0-9" years to "90-99", and "101+". The resulting collection contains 2000 prompts in the training set and 500 in the test set. Same samples were also used to prepare the regression-formatted prompts, in which LLMs were instructed to predict the exact age in years. Finally, we prepared equal numbers of pairwise prompts, for which LLMs were instructed to select which of the two samples was older. The training cohort for this format was based on pairs sampled from 2,907 unique DNAm profiles, while the test set contained 668 unique subjects. Cross-study comparisons were enabled for the pairwise prompts.

To represent the DNAm information in the prompts, we used beta-values from 1,000 CpG sites used by the DeepMAge aging clock (Fedor Galkin et al., 2020). For the single-profile formats, we selected 75 most and 75 least methylated sites, and for the pairwise prompts we showed 150 most dissimilar sites between two profiles. To reduce dependence on Illumina's internal site reference numbers, we referred to them by

genomic locations and probe target sequences instead, as described in Infinium Human Methylation platforms' manifest files. All coordinates were stated in the hg38 human genome assembly convention. For sites located near or within known genes, brief annotation of these genes were included, as per the platforms' manifest files.

Transcriptomic aging

We used open access data from GTEX, governed by the GTEx Portal Data License, to instruct the models to predict the age of multi-tissue human samples. We used 387 subjects for the training set and 162 for the test set. The data scope in GTEX contained 31 different tissues six age groups binned by decade (from "20-29" to "70-79").

The prompt collection include multiple choice binned age prediction, pairwise binary and ternary formats, as well a generative format. All formats contained 4,500 training and 1,200 test prompts. For the pairwise choice task, the binary format implied pairs of samples in which one sample was always older than the other, and for the ternary format, both samples could be from the same age group. No cross-tissue pairs were included in the pairwise prompt collections.

In the pairwise setting, top-100 expressed genes were presented for each sample and supplemented with contrastive ssGSEA finding, similar to the approach used in the cancer survival task. In the binned age prediction, 200 genes with numeric intensity values from each decile were shown, alongside the number of gene with nTPM<0.5 (marked as "not detected") and top-200 highly expressed genes. For the generative task, we presented LLMs with random 50 genes from among top-100 expressed genes in a sample and instructed the models to generate the remaining 50 in any order.

Proteomic aging

To assess if LLMs can pick up the aging signal in proteomic data, we used the PAD000022 Olink Explore 3072 blood plasma dataset uploaded to Pride as part of the work published in (Kirsher et al., 2025). The dataset features 78 subjects evenly distributed between young (18-22 years) and old (55-65 years) groups. We randomly assigned each 50% subject to the training or the test sets and prepared three prompt formats. In the binary classification format, LLMs were instructed to select whether a sample is old or young, based on the top-100 most expressed proteins in a profile with exact NPX values.

In the pairwise format, both sets contained 380 prompts, in which an LLM was instructed to select the older sample, based on the top-100 most differentially abundant proteins. In the generative format, a model was presented with randomly sampled 25 proteins among the top-50 most highly expressed and instructed to generate the remaining 25. To increase the number of generative prompts, we split profiles into parts representing Olink's categorization of sub-panels (Cardiometabolic, Inflammation, Neurology, and Oncology) and thus prepared four prompts from each sample. To account for a limited number of features measured by Olink platforms, we included all proteins a sub-panel could measure as part of the prompt. For all formats, control probes and probes below the limit of detection were excluded.

2.2. LLM benchmarking

We evaluated 15 LLMs across seven biomedical domains with tasks encompassing aging biology, survival analysis, and omics interpretation. The models tested included models developed by OpenAI (GPT-4.1, GPT-4.1-mini, GPT-5, GPT-5-mini, GPT-5.2, o3, o4-mini), Anthropic (Claude Sonnet 4.5), Google (Gemini 2.5 Pro, Gemini 2.5 Flash, Gemini 3 Pro Preview, Gemini 3 Flash Preview), xAI (Grok-3, Grok-4), and DeepSeek (DeepSeek-R1-0528).

Each task comprised multiple subtask formats designed to probe different aspects of model capability and assess the effect of different formats on model rankings. Classification tasks presented binary or multiple-choice questions, while pairwise comparison are a special case of a binary choice in which the labels represent not classes but entities, such as patient, individuals, or samples. Regression tasks required exact numerical predictions, while generative tasks required models to predict sets of genes characteristic of a given biological state (age and tissue). All prompts were constructed programmatically from source datasets and formatted with structured XML-like tags delineating the question, context, answer options, and ground truth. LLMs received task-specific system prompts instructing them to return structured JSON responses of an appropriate format. Models were queried through their respective APIs to Azure, Open Router, and xAI endpoints.

For unification purposes, the performance of LLMs in all binary, multiclass, and pairwise tasks is reported as unadjusted accuracy; in all generative tasks — as Jaccard index; in all regression tasks — as mean absolute error (MAE) in task-specific units.

3. Results

3.1. Benchmark and SFT dataset composition

We have aggregated data with various levels of annotation from seven disparate sources to compile a dataset that can both be used to test an LLM's grasp of aging biology as well as fine-tune public models for aging-related research tasks.

The totality of the resulting prompt collection contains 122,999 prompts, which roughly corresponds to 232MM tokens, when tokenized with OpenAI's tokenizer (see **Table 1**). In most cases, the number of prompts achieved has not exhausted all the information available in a source, meaning that an even more extensive collection may be compiled with these resources for SFT and benchmarking purposes.

Table 1. Breakdown of the LongevityBench prompt collection by data source and biological domain. The LongevityBench prompts were prepared using seven independent data sources to display aging through the lenses of different biodata types. All tasks support multiple prompt formats and annotation depths to control semantic and contextual confounders. Each task format is additionally split into train and test collection to enable SFT without data leaks.

Task category	Primary source	Description	N prompts	N tokens
General population survival	NHANES-IV	Given clinical blood tests and medical records of an individual, assess an individual's survival time.	4.93e+4	2.38e+7
Masked gene aging	Open Genes	Given a description of a gene and its effects in model organism, establish its expression trajectory in a human tissue.	2.95e+3	1.57e+6
Multi-mutant lifespan	Synergy Age	Given the effect of one mutation on lifespan in a model organism, assess the effect of an additional mutation relative to wild type.	4.72e+3	1.67e+6
Cancer survival	TCGA	Given a patient's diagnosis and tumor mRNA expression data, assess progression free survival.	3.46e+4	1.92e+7
Methylation aging	GEO	Given a person's blood DNA methylation levels, predict their age.	7.51e+3	1.41e+8
Transcriptomic aging	GTEX	Given a person's RNA expression levels, predict their age.	2.28e+4	4.37e+7
Proteomic aging	(Kirsher et al, 2025)	Given a person's Olink proteomics, predict their age.	1.16e+3	1.09e+6
TOTAL			1.23e+5	2.32e+8

3.2. Multi-task benchmarking

We evaluated 15 SotA LLMs across 17 LongevityBench tasks to establish a comprehensive ranking of their grasp of aging biology. The aggregate ranking, computed as the mean rank across all tasks, identified OpenAI's o3 as the top-performing model with an average rank of 4.65, followed by GPT-5 (5.06) and Gemini 3 Pro (5.59). Despite high average scores achieved by some models (**Table 2**), there is major heterogeneity in model performance across data modalities. As such, no single model achieved top-3 placement across all tasks, and even the highest-ranked models low results in certain domains (**Figure 2**).

The binary test results demonstrate that most models perform above the 0.5 random baseline, though margins vary considerably by task category. In the cancer survival task based on TCGA RNAseq data, Claude Sonnet 4.5 achieved the highest accuracy (0.697) when guessing which of the two patients had a longer PFS, with the close second being GPT-5.2 (0.651). For general population 10-year binary survival prediction based on NHANES clinical-level data, Gemini 3 Pro led with 0.882 accuracy, closely followed by o3 (0.881) and GPT-5 (0.874). Yet, in the pairwise setting using the same data (which individual lives longer?), all Gemini and xAI models failed to outperform random chance, while DeepSeek-R1, Sonnet-4.5, and OpenAI models showed practically indistinguishable results in the 0.64−0.65 range. The multi-mutant lifespan task, based on data from in murine and fly genetic experiments, had Gemini 3 Pro (0.781) and 2.5 Pro (0.734) taking the first two places, followed by Deepseek-R1 (0.717) (**Figure 3**).

The omics-level aging tasks presented greater challenges. In the transcriptomic pairwise age comparisons, we asked LLMs which of the two GTEx RNAseq profiles belonged to an older person, what we expected to be a relatively simple setting. However, all LLMs produced results indistinguishable from random baseline (0.48−0.54), suggesting that SotA LLMs struggle to extract aging signals from multi-tissue gene expression profiles. Notably, when the same data was presented as a multiclass age group prediction task (six decade bins covering the range of 20−79 years), LLM performance improved substantially: Gemini 2.5 Pro achieved 0.257 accuracy, well above the 0.167 random baseline. Furthermore, all developers except xAI had at least one model above the 0.20 accuracy threshold. This discrepancy indicates that aging-related transcriptomic patterns are encoded in these models, but the pairwise comparison format fails to elicit this knowledge. Proteomic aging prediction showed somewhat better discrimination in the pairwise format, with GPT-5 achieving 0.767 accuracy, though the binary classification (young vs old age groups) results for the same data type remained near random assignment level for most models with only GPT-5-mini, GPT-5.2, and GPT-5 showing accuracy above 0.55. In epigenetic aging, the pairwise format demonstrated that some SotA LLMs (Gemini and OpenAI families) can successfully detect aging signatures while others offered near-random predictions.

When all binary tasks are considered, models developed by OpenAI and Google dominate the leaderboard, with Gemini 3 Pro taking first place in 5 out of 10 tasks. Meanwhile, xAI models and Anthropic's Sonnet-4.5 generally show middling or low performance, with the notable exception of Sonnet's top-1 score in the cancer survival task (0.697 accuracy).

Since some tasks in LongevityBench are based on the same data sources, comparing LLMs based on the full array on 17 tasks may be unfair. To alleviate this bias toward models with a knack for overrepresented domains, we have selected one discriminant binary task per data source. Such tasks were selected to yield the greatest difference between minimal and maximal LLM scores. In this subset of tasks, Gemini-3 Pro shows outstanding performance compared to all others, but no single model dominates all domains. (**Figure 4**).

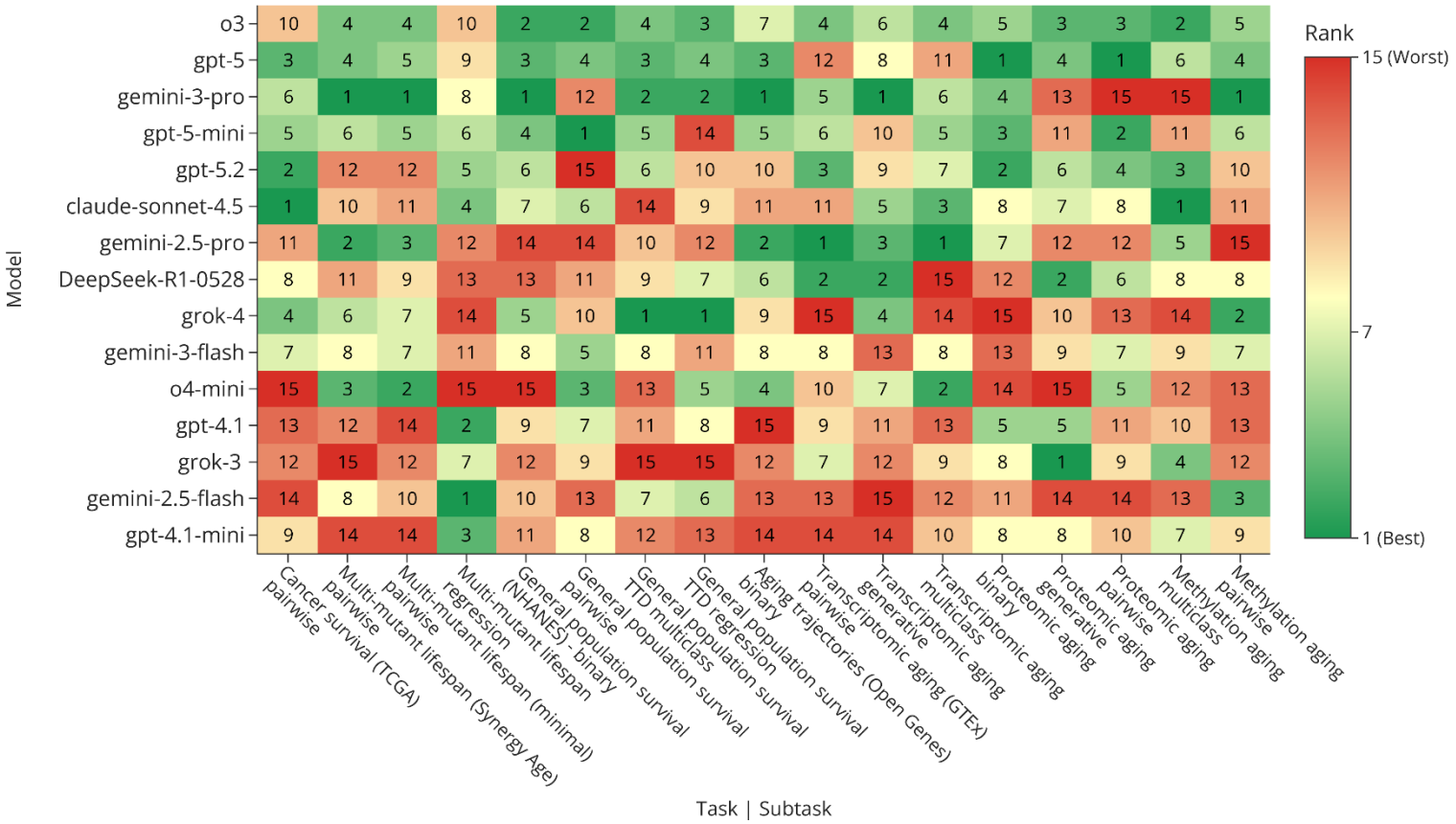

Figure 2. The ranks of 15 LLMs tested in 17 LongevityBench tasks. This collection of tasks represents common problems and settings encountered in aging research, such as survival analysis, omics-based age and lifespan prediction, determining the role of a gene in the aging process. The biodata domains covered by LongevityBench include: transcriptomics, epigenetics, proteomics, clinical blood tests, biometrics, and medical records.

Rows ordered by an LLM's average rank (see also **Table 2**).

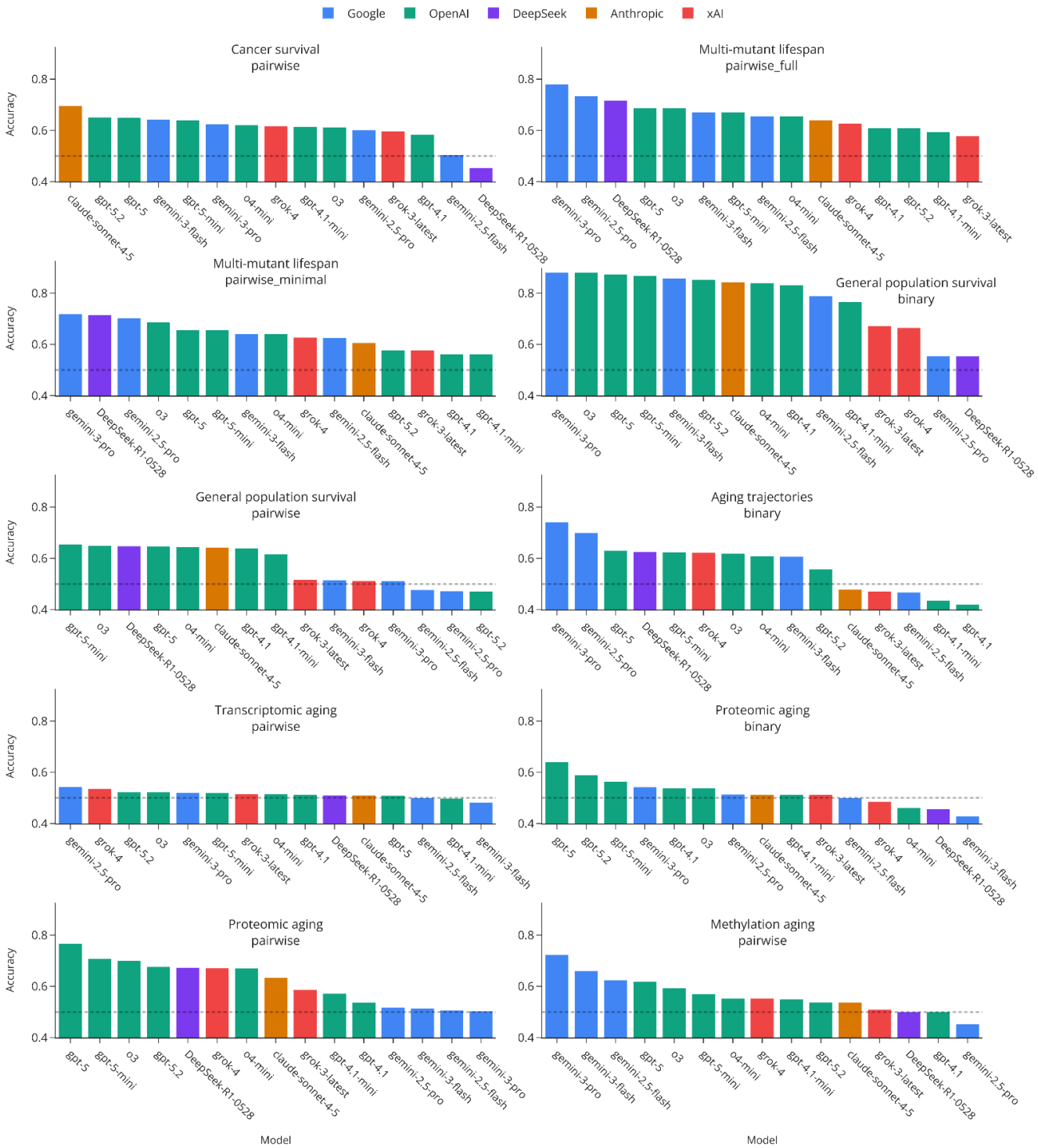

Figure 3. The performance of 15 tested LLMs in 10 tasks following a binary answer format. The presented tasks represent a variety problems encountered in aging research: survival analysis ("Cancer ..." and "General population survival"), omics-based age prediction ("Transcriptomic ...", "Proteomic ...", "Methylation aging"), establishing the effect of a gene on longevity ("Multi-mutant lifespan"), and detecting an aging-related expression pattern ("Aging trajectories"). See **Methods** for the exact definition of each task.

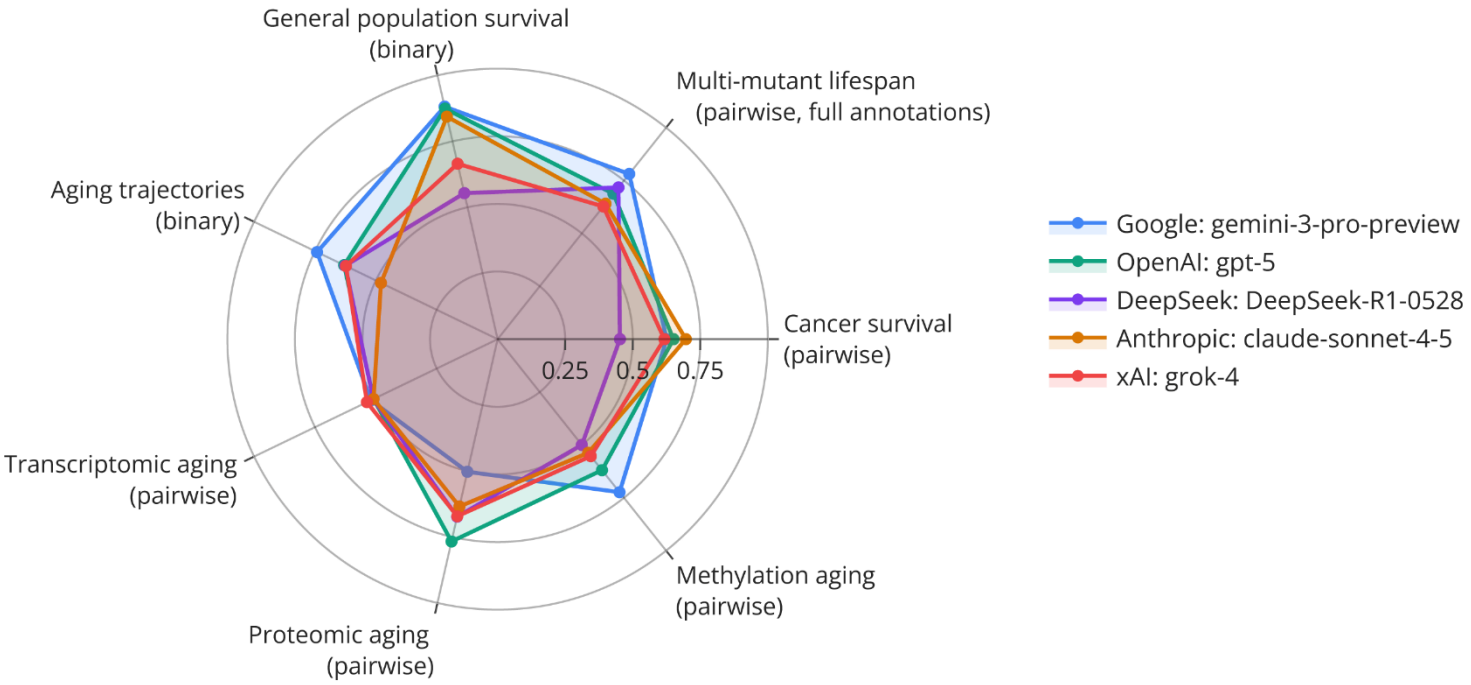

Figure 4. The comparison of best-performing LLMs from each developer across eight tasks selected for high inter-model variance. This tasks subset was defined to reduce the bias in model rankings stemming from the overrepresentation of certain data domains among the full set of 17 tasks.

Table 2. LongevityBench metrics across 17 tasks and 15 models.

* denotes the best-performing model in each task

Developer	Task	Cancer survival	Multi-mutant lifespan			General population survival				Aging trajec-tories
		Pairwise	Pair-wise (full)	Pair-wise (min)	Regres-sion	Binary	Pairwise	TTD multi-class	TTD regres-sion	Binary
	Metric	Accuracy	Accu-racy	Accu-racy	MAE, %	Accu-racy	Accu-racy	Accu-racy	MAE, months	Accuracy
Anthropic	**sonnet-4.5**	0.697 *	0.641	0.607	20.23	0.844	0.643	0.266	80.666	0.479
Deep Seek	**Deep Seek-R1**	0.453	0.717	0.714	20.37	0.554	0.648	0.274	71.829	0.626
Google	**gemini-2.5-flash**	0.504	0.656	0.625	19.07 *	0.789	0.478	0.319	72.714	0.468
	gemini-2.5-pro	0.602	0.734	0.703	27.35	0.555	0.473	0.295	84.749	0.700
	gemini-3-flash	0.643	0.672	0.641	28.34	0.858	0.515	0.389 *	55.378 *	0.608
	gemini-3-pro	0.625	0.781 *	0.719 *	22.63	0.882 *	0.512	0.386	64.903	0.742 *
OpenAI	**gpt-4.1**	0.584	0.609	0.562	19.62	0.832	0.639	0.277	80.161	0.420
	gpt-4.1-mini	0.614	0.594	0.562	20.03	0.766	0.617	0.276	85.614	0.436
	gpt-5	0.651	0.688	0.656	24.37	0.874	0.647	0.360	66.918	0.631
	gpt-5-mini	0.640	0.672	0.656	20.93	0.869	0.654 *	0.347	85.729	0.624
	gpt-5.2	0.651	0.609	0.578	20.41	0.853	0.471	0.322	81.260	0.558
	o3	0.613	0.688	0.688	25.79	0.881	0.650	0.347	66.278	0.620
	o4-mini	0.621	0.656	0.641	26.08	0.839	0.644	0.316	82.198	0.609
xAI	**grok-3**	0.598	0.578	0.578	21.74	0.672	0.518	0.228	89.346	0.471
	grok-4	0.617	0.627	0.627	27.96	0.665	0.513	0.315	75.092	0.623

Continued on the next page

Developer	Task	Transcriptomic aging			Proteomic aging			Methylation aging		Average Rank
		Gene-rative	Multi-class	Pair-wise	Binary	Gene-rative	Pair-wise	Multi-class	Pairwise	
	Metric	Jaccard	Accu-racy	Accu-racy	Accu-racy	Jac-card	Accur acy	Accu-racy	Accu-racy	
Anthropic	**sonnet-4.5**	0.139	0.234	0.51	0.513	0.012	0.634	0.210 *	0.537	**7.47**
Deep Seek	**Deep Seek-R1**	0.111	0.242	0.511	0.457	0.004	0.673	0.078	0.500	**8.41**
Google	**gemini-2.5-flash**	0.047	0.135	0.500	0.500	0.004	0.507	0.077	0.625	**10.41**
	gemini-2.5-pro	0.163	0.257 *	0.543 *	0.514	0.005	0.518	0.141	0.453	**8.06**
	gemini-3-flash	0.153	0.125	0.482	0.429	0.007	0.515	0.029	0.661	**8.53**
	gemini-3-pro	0.176 *	0.220	0.521	0.543	0.004	0.503	0.025	0.724 *	**5.59**
OpenAI	**gpt-4.1**	0.089	0.133	0.513	0.538	0.016	0.537	0.112	0.500	**9.88**
	gpt-4.1-mini	0.064	0.181	0.499	0.513	0.011	0.572	0.125	0.549	**10.47**
	gpt-5	0.103	0.175	0.510	0.641 *	0.017	0.767 *	0.137	0.620	**5.06**
	gpt-5-mini	0.090	0.223	0.520	0.564	0.006	0.708	0.107	0.570	**6.24**
	gpt-5.2	0.091	0.190	0.523	0.59	0.014	0.676	0.196	0.538	**7.24**
	o3	0.117	0.229	0.522	0.538	0.021	0.700	0.198	0.594	**4.65 ***
	o4-mini	0.072	0.183	0.515	0.462	0.010	0.671	0.116	0.553	**8.71**
xAI	**grok-3**	0.082	0.182	0.516	0.513	0.030 *	0.587	0.146	0.510	**10.12**
	grok-4	0.174	0.104	0.537	0.486	0.022	0.672	0.117	0.553	**8.41**

3.3. Format-dependent variations

While using LongevityBench, we noticed that model rankings were not stable across question formats. We first examined this phenomenon using the multi-mutant lifespan task, where prompts were prepared in both minimal and extended annotation formats. The minimal format provided only the lifespan effects of individual mutations, while the extended format included GO annotations and experimental context from source publications. This comparison tests whether models can leverage additional biological context to improve phenotype-level predictions.

Several models demonstrated substantial improvement when provided with richer annotations: Gemini 3 Pro improved from 0.719 to 0.781, Gemini 2.5 Pro from 0.703 to 0.734. Similarly, GPT-4.1's accuracy improved from 0.562 to 0.609, but the model with the best average rank, o3, remained stable at 0.688 in both conditions, suggesting that this LLM might have internalized to all the additional information we presented at training. Both tested xAI and DeepSeek-R1models showed no significant improvement as well. These patterns suggest that the capacity to extrapolate biological annotations into high-level phenotypes, such as longevity, varies across model families (**Figure 5**).

The multiclass task formulations revealed additional format-dependent effects. In the NHANES survival prediction, the binned TTD format (four categories: 0−5, 5−10, 10−15, 15+ years) yielded different model rankings than the binary 10-year survival task using the same underlying data. Gemini 3 Flash achieved the highest multiclass accuracy (0.389), followed by Gemini 3 Pro (0.386), yet these same models ranked below random chance in the pairwise comparison format. Similarly, in methylation-based age prediction, Claude Sonnet 4.5 achieved the highest multiclass accuracy (0.211) despite middling performance in the pairwise setting (**Figure 6**).

The regression tasks revealed systematic biases in model predictions. For multi-mutant lifespan effect prediction, models achieved MAE values ranging from 19.1% (Gemini 2.5 Flash) to 28.3% (Gemini 3 Flash), with most models showing reasonable regression parameters (**Figure 7**). However, the general population survival regression task exposed systematic underestimation: all models compressed their predictions toward the 50−100 month range, while true survival times extended beyond 250 months. This stands in stark contrast to the strong binary classification performance most models achieved on the same NHANES data, where accuracies exceeded 0.85 for top performers. The best regression model (Gemini 3 Flash, MAE 55.4 months) still exhibited this compression pattern, and weaker models (Grok 3 Latest, MAE 89.3 months) showed severe underestimation of long-term survival (**Figure 8**). The generative tasks, which required models to predict highly expressed genes given partial expression profiles and sample's age and tissue metadata,

revealed pronounced disparity between omics modalities. For transcriptomic aging, the top models achieved Jaccard similarities of 0.176 (Gemini 3 Pro) and 0.174 (Grok 4), indicating meaningful overlap with ground-truth gene sets. On the other hand, proteomic aging generation proved far more challenging despite a much narrower background set of measurable proteins. The best Jaccard similarity was only 0.030 (Grok 3 Latest), representing a six-fold reduction compared to transcriptomic performance (**Figure 9**).

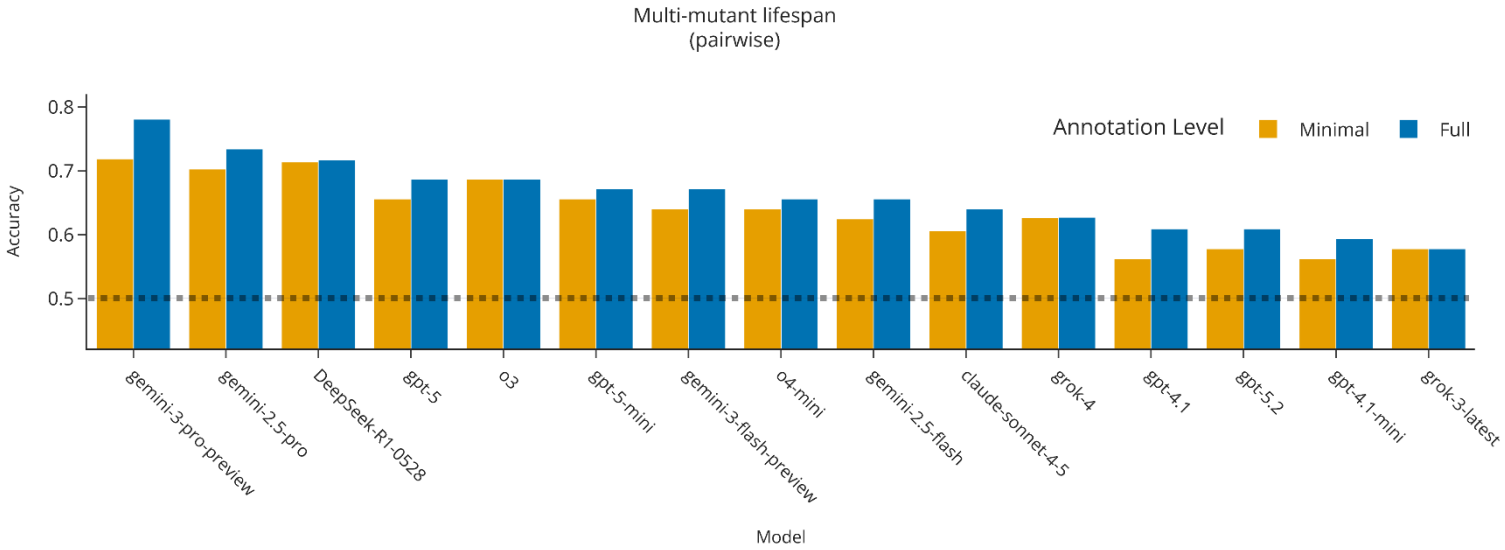

Figure 5. The change in LLMs' performance in the "Multi-mutant lifespan" task caused by adding extra context to test prompts. Some models determine the longer-living genotype in *D.melanogaster* and *M.musculus* more accurately when presented with experiment protocols and descriptions of the affected genes.

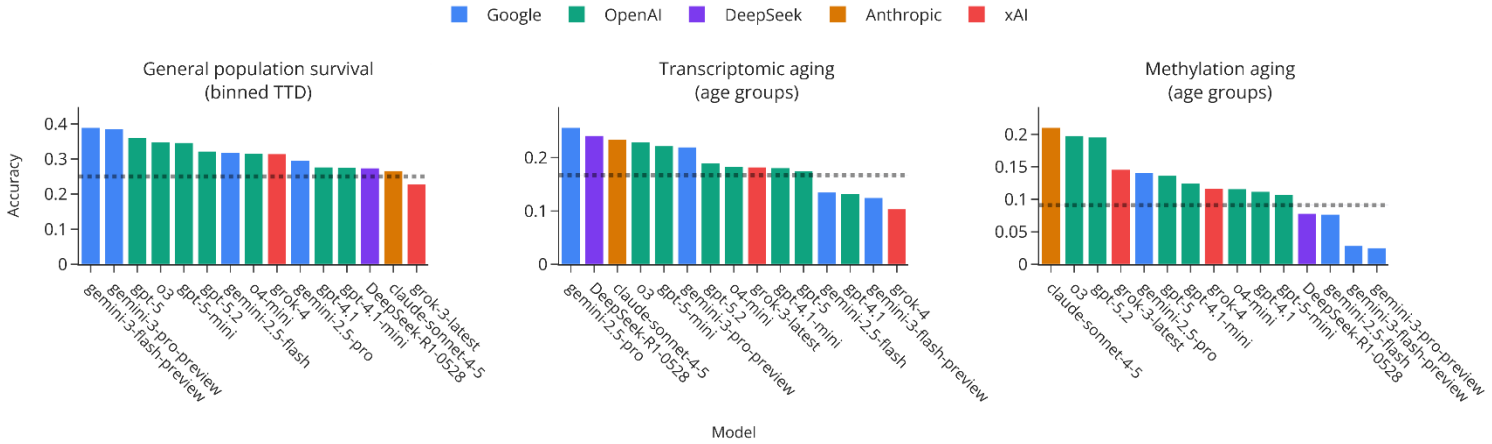

Figure 6. LLM performance in multiclass tasks from LongevityBench. Changing the format of the answer from pairwise choice (which sample is older or lived longer?) to multiclass age group prediction significantly affects LLM ranking (compare to **Figure 3**)

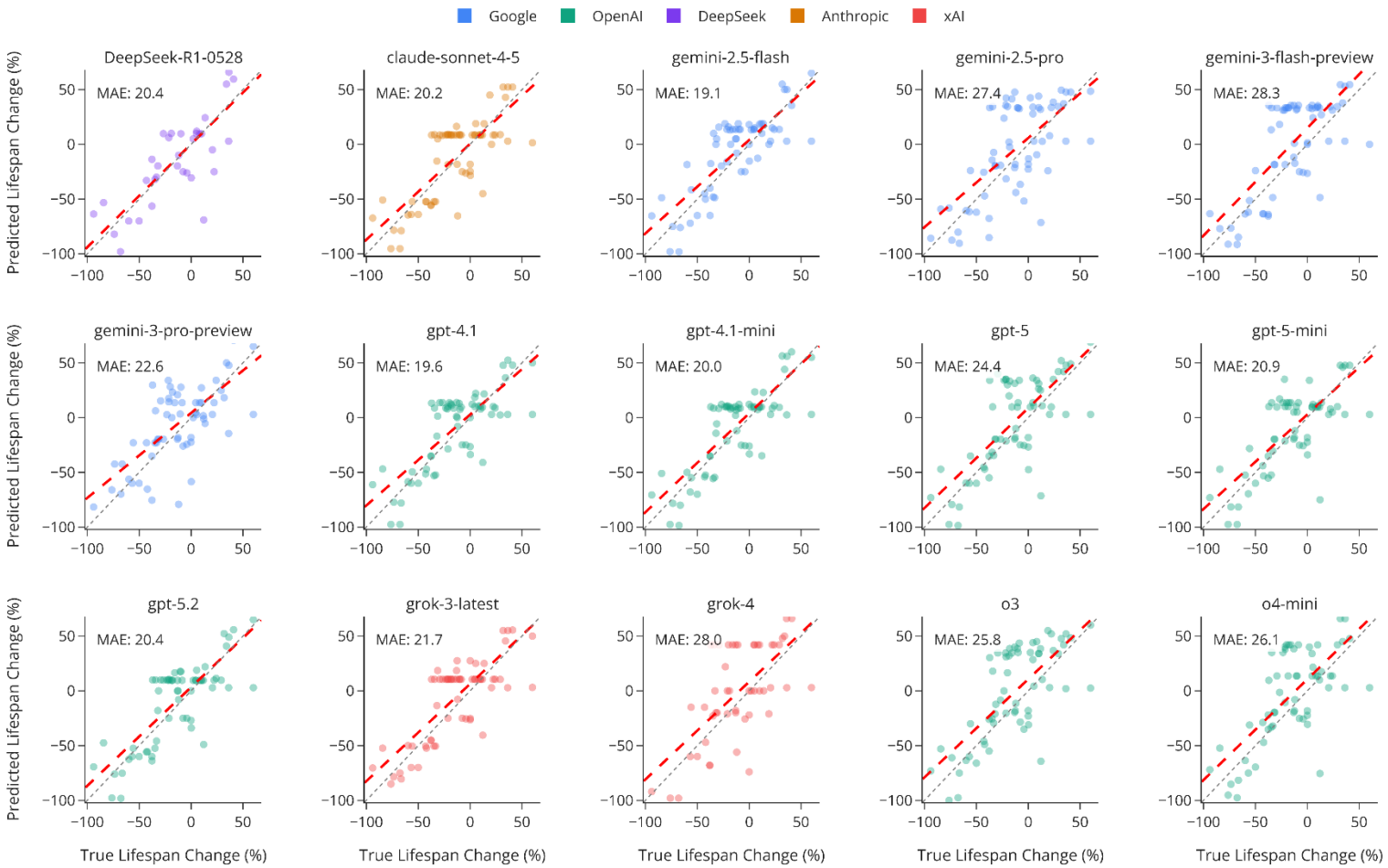

Figure 7. The performance of 15 LLMs in the "Multi-mutant lifespan" task which requires the models to provide the exact change in lifespan relative to wild type. The LLMs predict the combined effect of two mutations on model organisms' lifespan with a 19-28% mean absolute error (MAE).
Red line represents the ordinary least squares regression.

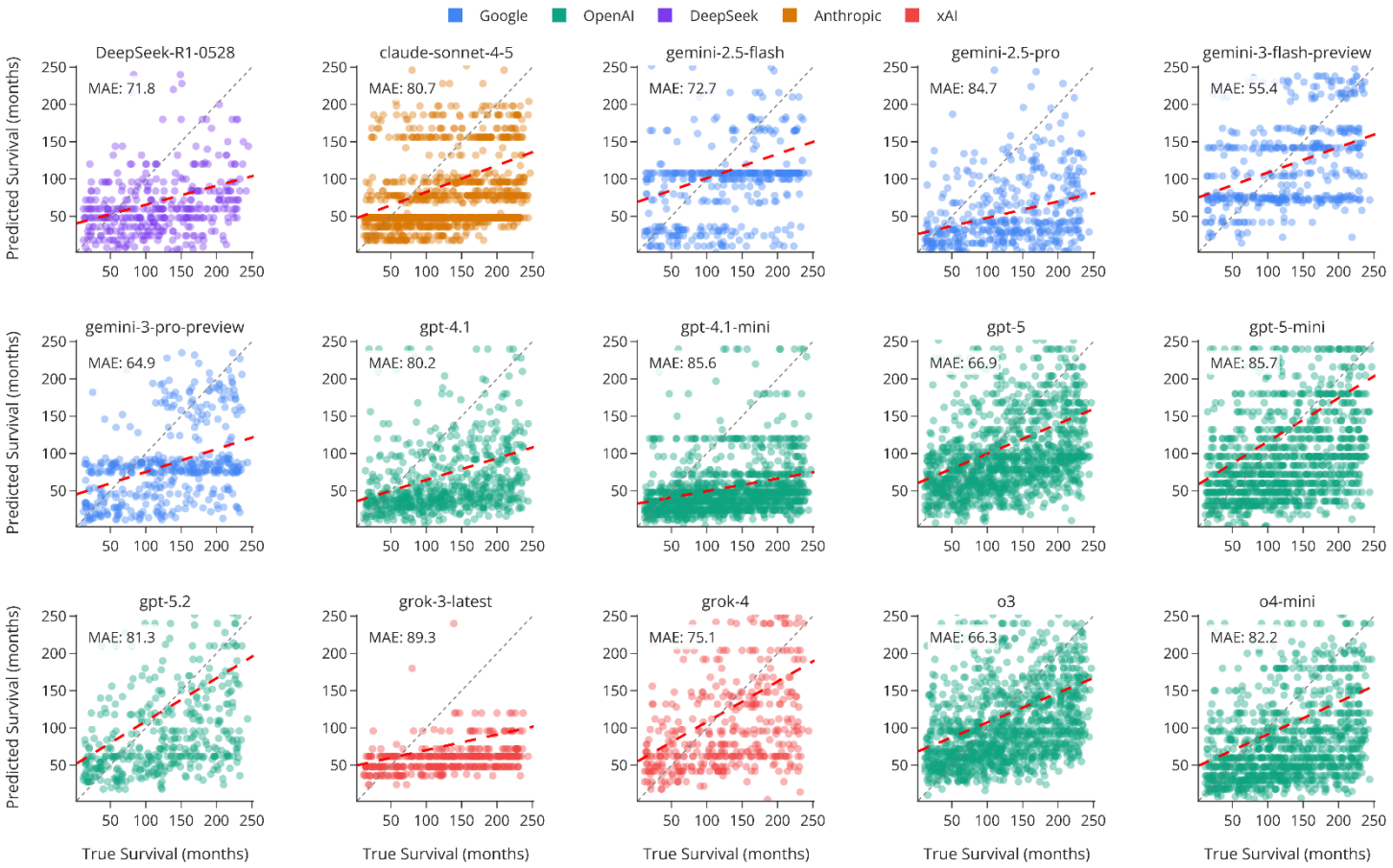

Figure 8. The performance of 15 LLMs in the "General population survival" task which requires the models to present the exact time-to-death (TTD) of an individual.
Most LLMs tend to underestimate the TTD based on medical history and clinical blood tests, as indicated by the greater density of dots below the unity line (grey).
Red line represents the ordinary least squares regression.

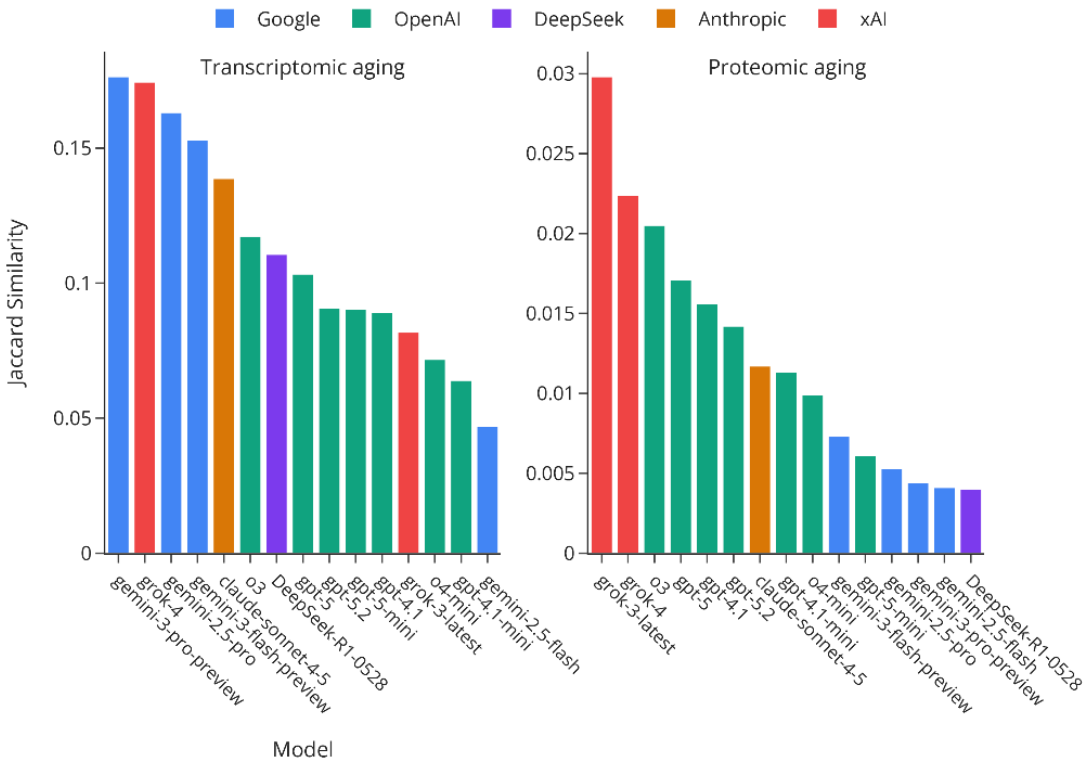

Figure 9. LLMs' performance in generative omics tasks. Some LLMs can somewhat accurately complete lists of highly expressed genes in GTEx samples (Jaccard index > 0.15) but no LLM can do the same with Olink data (Jaccard index < 0.03) despite a much smaller background set of measurable genes.

4. Discussion

Our systematic evaluation of 15 SotA LLMs using LongevityBench reveals that no current model is reliable enough to serve as a competent aging research companion. While individual models achieved strong performance on specific tasks, the substantial cross-task variance demonstrates that success in one domain does not guarantee adequate responses in others (**Figures 2-4**). Researchers selecting LLMs for aging-related work must therefore consider specific task requirements rather than relying on general capability benchmarks or aggregate rankings.

The instability of model rankings across question formats represents a methodological concern for AI-assisted research. When the same NHANES survival data was presented in binary, pairwise, multiclass, and regression formats, model rankings shifted dramatically: Gemini 3 Pro achieved 0.882 accuracy in binary classification but fell below random chance in pairwise comparisons. Similarly, transcriptomic age prediction yielded near-random

pairwise accuracies (0.48−0.54) yet meaningful multiclass performance (up to 0.257 for Gemini 2.5 Pro for six age groups, **Figure 6**). These observations imply that benchmark design choices greatly influence conclusions about model capabilities, and that reported metrics may not generalize to reformulated queries addressing an otherwise familiar LLM domain.

The systematic underestimation of survival times in NHANES regression warrants particular attention (**Figure 8**). All tested models compressed predictions toward 50–100 months, despite achieving strong binary classification accuracy (>0.85) on the same data. This bias likely reflects training corpora where clinical risk factors co-occur with mortality-related text. Models appear to treat any mention of serious disease as a substantial mortality risk, failing to appropriately weight information indicating successful treatment or long-term remission. This hypothesis, however, requires further investigation.

The six-fold performance gap between transcriptomic and proteomic generative tasks (Jaccard 0.176 vs 0.030, **Figure 9**) suggests that LLMs have internalized mRNA expression patterns more effectively than protein abundance profiles. Several factors may contribute: the relative novelty of platforms such as Olink Explore 3072, underrepresentation of proteomics in training corpora, and potentially false assumptions about protein-RNA co-expression derived from publications reporting primarily transcriptome findings. This cross-omic disparity highlights the risks of applying LLMs to emerging data modalities without dedicated evaluation.

Among tested models, the Gemini family demonstrated the most consistent performance across task types with Gemini-3 Pro ranking first in 6 out of 17 tasks (**Figures 2,4**). Conversely, DeepSeek-R1-0528 ranked last most frequently (4 tasks), yet still occasionally reaching the top-3 position, indicating domain-specific strengths. Reasoning-focused o3 LLM is recognized as the most aging-aligned model, according to the average rank across all tasks. However, o3 did not show outstanding performance and did not take first place in any of the tests. Moreover, its long generation time, a result of its reasoning-focused design, may limit o3's utility in some high-throughput applications.

We intend LongevityBench to serve as both a selection criterion for existing models and as an optimization target for specialized training. By establishing quantitative baselines across diverse aging-related tasks, this benchmark provides an essential sub-module of the MMAI Gym for Science. The variety of tasks included in LongevityBench makes it a perfect toolset for enabling genuine biological understanding within LLMs.

5. Limitations

This benchmark did not assess inference costs or latency, which remain relevant for high-throughput research pipelines. Some models require 30–50 seconds per response, and reasoning models may consume substantially more tokens, creating barriers to large-scale deployment regardless of accuracy.

We note that model compliance with output format instructions varied, and failures to produce parseable responses were treated as missing data. While this approach avoids penalizing models for refusals, it obscures practical considerations: models that frequently deviate from specified formats complicate automation efforts. We aim to cover the metrics of execution time, costs, and format compliance rates in later iterations of LongevityBench.

The scope of tasks presented in this work does not cover all the settings encountered in aging research. We plan to continue growing the array of tests shown here to next include data from chemical screening and geroprotector experiments, cross-species translation, sequence-based, and other tasks.

6. Resource availability

Current and historical leaderboard standings for all tasks and LLMs are available at bench.insilico.com.

7. Supplementary materials

Prompt samples

General population survival

Binary

```
<question>
      Did this participant survive 10 years post-examination?
</question>
<options>
      A. Yes  B. No
</options>
<annotation>
      <demographic>
```

```
            67-year-old male with college graduate or higher education of non-
            Hispanic white ethnicity rates health as excellent.
      </demographic>
      <biometric>
            BMI 25.1 kg/m2, height 177.7 cm, waist circumference 97.7 cm, body
            weight 79.3 kg, diastolic blood pressure 79.3 mm Hg, systolic
            blood pressure 130.7 mm Hg, pulse 51.3 bpm.
      </biometric>
      <blood>
            Albumin (g/dL) 4.30; Alkaline phosphatase (U/L) 94.00; Blood urea
            nitrogen (mg/dL) 11.00; …[TRUNCATED]… Vitamin B12 (pmol/L) 261.25.
      </blood>
      <questionnaire>
            Rates own health as excellent. Family history of diabetes.
      </questionnaire>
</annotation>
<answer>
            A
</answer>
<follow_up>
      SEQN 11005: Alive at 18 years 5 months (censored)
</follow_up>
```

Pairwise

```
<question>
      Which participant had a longer survival period after examination?
</question>
<options>
      A. Participant A  B. Participant B
</options>
<participant_a>
      <demographic>
            76-year-old male with less than 9th grade school education of non-
            Hispanic black ethnicity.
      </demographic>
      <biometric>
```

BMI 29.1 kg/m2, height 160.5 cm, body weight 75.0 kg, diastolic blood pressure 60.0 mm Hg, systolic blood pressure 122.0 mm Hg, pulse 62.0 bpm.
</biometric>
<blood>
Albumin (g/dL) 3.90; ...[TRUNCATED]... Vitamin B12 (pmol/L) 580.07.
</blood>
<medication>
Donepezil (taken for 4 years), Haloperidol (taken for 10 months).
</medication>
<questionnaire>
Has medical history of arthritis; cancer.
</questionnaire>
</participant_a>
<participant_b>
<demographic>
80-year-old male with college graduate or higher of non-Hispanic white ethnicity.
</demographic>
<biometric>
BMI 30.7 kg/m2, height 171.8 cm, waist circumference 109.2 cm, body weight 90.7 kg, diastolic blood pressure 72.7 mm Hg, systolic blood pressure 121.3 mm Hg, pulse 48.7 bpm.
</biometric>
<blood>
Albumin (g/dL) 4.50; ...[TRUNCATED]... Vitamin B12 (pmol/L) 333.58.
</blood>
<medication>
Celecoxib (taken for 2 weeks).
</medication>
<questionnaire>
Has medical history of asthma. Family history of diabetes, arthritis. Engages in moderate home or yard tasks.
</questionnaire>
</participant_b>

<answer>
A
</answer>
<follow_up>
Participant A (SEQN 18188): Deceased at 4 years 2 months, cause: Alzheimer's disease. Participant B (SEQN 788): Deceased at 1 years 2 months, cause: a heart condition
</follow_up>

TTD multiclass

<question>
How long did this participant survive post-examination?
</question>
<options>
A. 15+ years B. 10-15 years C. 5-10 years D. 0-5 years
</options>
<annotation>
<demographic>
80-year-old male with less than 9th grade of school education.
</demographic>
<biometric>
BMI 24.2 kg/m2, height 165.3 cm, waist circumference 103.0 cm, body weight 66.2 kg, diastolic blood pressure 88.0 mm Hg, systolic blood pressure 170.0 mm Hg, pulse 82.0 bpm.
</biometric>
<blood>
Albumin (g/dL) 4.30; ...[TRUNCATED]... Vitamin B12 (pmol/L) 168.26.
</blood>
<medication>
Beclomethasone (taken for 2 months), ...[TRUNCATED]...
</medication>
<questionnaire>
Rates own health as good. Has medical history of angina pectoris; stroke; curent asthma. Family history of heart attack.
</questionnaire>
</annotation>
<answer>

```
      D
</answer>
<follow_up>
      SEQN 12418: Deceased at 1 years, cause: a heart condition
</follow_up>
```

Aging trajectories

Binary

```
<question>
      You are presented with a description of an unknown gene's ("the gene")
      function and experimental findings in model organisms and humans.
      In CD8+ T cells of humans between ages 23 and 65 years, the gene's mRNA
      expression measured by microarray:
</question>
<options>
      A. increases with age B. decreases with age
</options>
<model_organisms>
      In the brain of rhesus monkeys between ages 5 and 31 years, the gene
      decreased expression by 43.0%, as established by measuring protein
      levels. ...[TRUNCATED]...
</model_organisms>
<ppi>
      This gene interacts with CHML, CHML, KMO, and KMO in humans.
</ppi>
<go>
      This gene has molecular functions including 11-cis retinal binding,
      ...[TRUNCATED]...
</go>
<protein_atlas>
      This gene mRNA (Tau=0.81, tissue-enhanced) is most expressed in cerebral
      cortex and liver (nTPM 6-11). ...[TRUNCATED]...
</protein_atlas>
<answer>
      B
</answer>
<follow_up>
```

Gene: OPN3. OPN3 showed decreased gene expression by 46.2% in CD8+ T cells between ages 23 and 65 years, measured by microarray. DOI: 10.1111/j.1474-9726.2009.00534.x
</follow_up>

Multi-mutant lifespan

Regression

<question>
What is the effect of the Cat(IF);Sod1(IF) genetic perturbation on the animal's lifespan in % when compared to wild type?
</question>
<organism>
Drosophila melanogaster
</organism>
<genes>
Cat: Cat (Catalase) | Sod1: Sod1 (Superoxide dismutase 1)
</genes>
<go>
Cat: Function: metal ion binding, oxidoreductase activity, ...[TRUNCATED]... | Sod1: ...[TRUNCATED]... Location: mitochondrial intermembrane space, peroxisome, cytoplasm
</go>
<experimental_context>
Temperature: 25°C Diet: cornmeal-agar medium ...[TRUNCATED]... Yeast FLP recombinase was used to allow induced overexpression of catalase and/or Cu/Zn-superoxide dismutase in adults. Expression of FLP recombinase was driven by the heat-inducible hsp70 promoter ...[TRUNCATED]...
</experimental_context>
<known_effects>
Cat(IF): -14.2% lifespan (39 days vs 45 days WT) | Sod1(IF): +27.8% lifespan (60 days vs 47 days WT)
</known_effects>
<answer>
+22.2%
</answer>
<follow_up>

Cat(IF): 39 days (-14.2% vs 45 WT) [PMID:9858546] | Sod1(IF): 60 days (+27.8% vs 47 WT) [PMID:9858546] | Cat(IF);Sod1(IF): 57 days (+22.2% vs 47 WT) [PMID:9858546] | Interaction: Opposite lifespan effects of single mutants
</follow_up>

Pairwise (full annotations)

<question>
Based on the known effects of individual interventions, predict which of them two interventions will result in longer mean lifespan. Consider potential genetic interactions.
</question>
<options>
A. Cat(IF);Sod1(IF) B. Cat(IF)
</options>
<organism>
Drosophila melanogaster
</organism>
<genes>
Cat: Cat (Catalase) | Sod1: Sod1 (Superoxide dismutase 1)
</genes>
<go>
Cat: Function: metal ion binding, ...[TRUNCATED]...| Sod1: Function: protein homodimerization activity, ...[TRUNCATED]...
</go>
<experimental_context>
Temperature: 25°C Diet: cornmeal-agar medium Fly stocks ...[TRUNCATED]...
</experimental_context>
<known_effects>
Cat(IF): -14.2% lifespan (39 days vs 45 days WT) | Sod1(IF): +27.8% lifespan (60 days vs 47 days WT)
</known_effects>
<answer>
A
</answer>
<follow_up>

```
      ...[TRUNCATED]...
</follow_up>
```

Cancer survival

Pairwise survival

```
<question>
      Which of the following two Melanoma cases has had a longer progression-
      free interval after the initial RNAseq screening:
</question>
<options>
      Patient-A: A 42-year-old female diagnosed with Stage IIIB ulcerated
      tumor (TT4b, NN1, MM0);
      Patient-B: A 77-year-old male diagnosed with Stage IIIC ulcerated tumor
      (TT4b, NN2b, MM0), Clark level IV
</options>
<gsea>
      Patient-A shows high activity in "Formation Of The Cornified Envelope"
      pathway (NES: 0.44, genes: KRT19, DSG2, KRT7, KRT18, KRT8 and 124 other
      genes), while Patient-B shows low activity (NES: -0.13, genes: PRSS8,
      DSG2, EVPL, PPL, DSP and 124 other genes). ...[TRUNCATED]... Both
      patients show suppressed activity in the "Defective C1Galt1C1 Causes
      Tnps" pathway, with Patient-B demonstrating particularly strong
      enrichment (NES: -0.30, genes: MUC4, MUC15, MUC13, MUC20, MUC6 and 12
      other genes) compared to the other patient (NES: -0.15, genes: MUC13,
      MUC16, MUC6, C1GALT1C1, MUC17 and 12 other genes).
</gsea>
<answer>
      Patient-B
</answer>
<follow_up>
      Patient-A: TCGA-EB-A4IQ-01: Progressed after 11.1 months;  Patient-B:
      (TCGA-GN-A26C-01: Progressed after 1.6 years)
</follow_up>
```

Methylation aging

Multiclass

<question>
Based on the following DNA methylation profile from a blood sample, which age group does this individual most likely belong to?
</question>
<patient>
A female from Denmark.
</patient>
<methylation_profile>
Lowest methylated (75 sites): (chr12:121078658, + strand, CTTCAGCAGCGACGTCTTGGACTCGCTTTTGGCGATGTGCGAGCTCATCG): Located in CABP1 (in first exon) (calcium binding protein 1 isoform 1, ...[TRUNCATED]... enzyme inhibitor activity). Also known as: CALBRAIN; HCALB_BR Beta: 0.01. | ...[TRUNCATED]...
||
Highest methylated (75 sites): (chr17:7833237, + strand, AATATTCAAGCTGCAGAGCCTTCAGCAGGCGGGGCTTATGCAGATGAGCG): Located in KCNAB3 (within 1500bp of TSS) (potassium voltage-gated channel; ...[TRUNCATED]... potassium ion transport). Also known as: AKR6A9; KCNA3B; KCNA3.1B; KV-BETA-3; MGC116886 Unclassified Cell type specific. Beta: 0.86. | ...[TRUNCATED]...
</methylation_profile>
<options>
A. 0-9 B. 60-69 C. 20-29 D. 100+ E. 80-89 F. 50-59 G. 90-99 H. 70-79 I. 40-49 J. 10-19 K. 30-39
</options>
<answer>
B
</answer>
<follow_up>
True age: 62 years, age group 60-69 (GSM1506320, GSE61496).
</follow_up>

Pairwise

<question>
Based on the following DNA methylation profiles from blood samples, which individual is older?
</question>
<options>

A: A female from Estonia ; B: A female from United Kingdom, diagnosed with nephropathy and type 1 diabetes.
</options>
<methylation_data>
The following CpG sites show the largest methylation differences between the two individuals: (chr1:247712591, + strand, CGCATTTGTAATGGTGAGGTACATTGTGCTTCTAGAAGGTTAAGCCTGAA): Located in C1orf150 (in gene body) (hypothetical protein LOC148823, synonyms: FLJ44728, RP11-978I15.8). Also known as: FLJ44728; RP11-978I15.8 Patient-A: 0.74, Patient-B: 0.19. | ...[TRUNCATED]...
</methylation_data>
<answer>
B
</answer>
<follow_up>
Patient-A: 25 years (GSM1425751, GSE59065). Patient-B: 39 years (GSM501660, GSE20067).
</follow_up>

Transcriptomic aging

Pairwise

<question>
Which of the following two Skin samples comes from an older individual?
</question>
<options>
Sample-A: male, top expressed genes: KRT1; KRT10; COX1; ND4; ...[TRUNCATED]...
Sample-B: male, top expressed genes: FN1; COX1; COL1A2; ND4; ...[TRUNCATED]...
</options>
<gsea>
Sample-A shows moderate activity in Formation Of The Cornified Envelope (NES: 0.18, genes: SPINK9, LCE4A, LCE3C, LCE3B, LCE3A and 124 other genes), while Sample-B shows very low activity (NES: -0.14, genes: LCE3A, KRT84, KRT40, LCE1D, LCE1E and 124 other genes). ...[TRUNCATED]... Both patients show suppressed activity in Fatty Acids, with Sample-B demonstrating particularly strong enrichment (NES: -0.31, genes: CYP2A13, CYP2B6, CYP2F1, CYP4A11, CYP4F2 and 10 other genes) compared

to the other patient (NES: -0.07, genes: CYP2F1, CYP2A13, CYP4F11, CYP2A7, CYP2B6 and 10 other genes).
</gsea>
<answer>
A
</answer>
<follow_up>
Sample-A (GTEX-148VJ-0626-SM-5LUAW): 70-79; Sample-B (GTEX-11WQK-0008-SM-5SI6T): 50-59
</follow_up>

Multiclass

<question>
Based on this expression profile, predict the chronological age group of the individual.
</question>
<options>
A. 70-79; B. 40-49; C. 50-59; D. 30-39; E. 20-29; F. 60-69
</options>
<context>
Human Brain transcriptomics data (RNA-seq, log1p-transformed TPM). Sex: male.
</context>
<expression>
3388 genes: not detected |
1076 genes (0.69-4.67): CSN1S1, DEFA8P, ...[TRUNCATED]... |
1079 genes (13.21-14.38): ANAPC5, CYC1 ...[TRUNCATED]... |
Top-200 expressed: ND4 (19.29), ND2 (19.20), ...[TRUNCATED]...
</expression>
<answer>
B
</answer>
<follow_up>
Sample ID: GTEX-ZUA1-0011-R8b-SM-51MST; Actual age: 40-49
</follow_up>

Generative

```
<instructions>
      Given the sample metadata and 100 highly expressed genes, predict 100
      additional highly expressed genes from the same sample. Output HGNC gene
      symbols separated by semicolons (;). Output exactly 100 genes.
</instructions>
<context>
      Tissue: Brain; Sex: female; Age group: 40-49
</context>
<input_genes>
      COX1; COX2; COX3; CYTB; SNAP25; ...[TRUNCATED]...
</input_genes>
<answer>
      ND4; ATP6; ND2; ND1; CALM1; PKD1; ...[TRUNCATED]...
</answer>
<follow_up>
      Sample ID: GTEX-12ZZX-0011-R11a-SM-5DUVJ; Age: 40-49
</follow_up>
```

Proteomic aging

Generative

```
<platform>
      Olink Explore 3072 proximity extension assay (PEA) plasma proteomics.
      Values are NPX (Normalized Protein eXpression, log2 scale).
      Cardiometabolic panel: Proteins pertinent to cardiovascular and
      metabolic disease biology, providing broad coverage of relevant pathways
      and processes. Includes both well-established biomarkers and more
      exploratory potential markers. | 729 measured proteins: AAMDC;ABCA2;
      ...[TRUNCATED]...
</platform>
<question>
      This plasma proteomics sample is from a 55–65 year old M. Given 25 of
      the top-50 most abundant proteins (NPX), predict the remaining 25 (in
      any order).
</question>
<olink>
      GLO1 (5.70); KIF22 (4.63); RANBP1 (4.17); EHD3 (3.68); ...[TRUNCATED]...
</olink>
<answer>
      CSRP3;MYH9;TNNI3;...[TRUNCATED]...
<follow_up>
      Subject: 5530211349; Age: 55–65; Sex: M
</follow_up>
```

Pairwise

```
<question>
      Based on their plasma proteome profiles, which individual is older?
</question>
<options>
      A. Patient-A B. Patient-B
</options>
<platform>
      Olink Explore 3072 proximity extension assay (PEA) plasma proteomics.
      Values are NPX (Normalized Protein eXpression, log2 scale). 2916
      proteins measured across Cardiometabolic, Inflammation, Oncology, and
      Neurology panels.
```

```
</platform>
<olink>
      Top-100 differentially abundant proteins (NPX): CASP1 (A: -8.89, B: -
      0.74); FOLR3 (A: 3.61, B: -4.33); ...[TRUNCATED]...
</olink>
<answer>
      B
</answer>
<follow_up>
      Patient-A: 5530211325, 18-22, F; Patient-B: 5530211357, 55-65, M
</follow_up>
```

Binary

```
<platform>
      Olink Explore 3072 proximity extension assay (PEA) plasma proteomics.
      Values are NPX (Normalized Protein eXpression, log2 scale). 2916
      proteins measured across Cardiometabolic, Inflammation, Oncology, and
      Neurology panels.
</platform>
<question>
      Based on the plasma proteome profile of this male individual, what is
      their age group?
</question>
<olink>
      Top-100 most abundant proteins (NPX): GLO1 (5.67); BANK1 (5.07);
      ...[TRUNCATED]...
</olink>
<options>
      A. 18-22 B. 55-65
</options>
<answer>
      B
</answer>
<follow_up>
      Subject: 5530211502; Age: 55-65; Sex: M
</follow_up>
```

8. Bibliography

Aliper, A., Zavoronkovs, A., Zhebrak, A., Kadurin, A., Polykovskiy, D., Shayakhmetov, R., 2022. Mutual information adversarial autoencoder [Patent]. US11403521B2.

Bobrov, E., Georgievskaya, A., Kiselev, K., Sevastopolsky, A., Zhavoronkov, A., Gurov, S., Rudakov, K., del Pilar Bonilla Tobar, M., Jaspers, S., Clemann, S., 2018. PhotoAgeClock: deep learning algorithms for development of non-invasive visual biomarkers of aging. Aging 10, 3249–3259. https://doi.org/10.18632/aging.101629

Bunu, G., Toren, D., Ion, C.-F., Barardo, D., Sârghie, L., Grigore, L.G., de Magalhães, J.P., Fraifeld, V.E., Tacutu, R., 2020. SynergyAge, a curated database for synergistic and antagonistic interactions of longevity-associated genes. Sci. Data 7, 366. https://doi.org/10.1038/s41597-020-00710-z

Fedor Galkin, Polina Mamoshina, Kirill Kochetov, Denis Sidorenko, Alex Zhavoronkov, 2020. DeepMAge: A Methylation Aging Clock Developed with Deep Learning. Aging Dis. 0. https://doi.org/10.14336/AD.2020.1202

Galkin, F., Mamoshina, P., Aliper, A., Putin, E., Moskalev, V., Gladyshev, V.N., Zhavoronkov, A., 2020. Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning. iScience 23, 101199. https://doi.org/10.1016/j.isci.2020.101199

Galkin, F., Naumov, V., Pushkov, S., Sidorenko, D., Urban, A., Zagirova, D., Alawi, K.M., Aliper, A., Gumerov, R., Kalashnikov, A., Mukba, S., Pogorelskaya, A., Ren, F., Shneyderman, A., Tang, Q., Xiao, D., Tyshkovskiy, A., Ying, K., Gladyshev, V.N., Zhavoronkov, A., 2024. Precious3GPT: Multimodal Multi-Species Multi-Omics Multi-Tissue Transformer for Aging Research and Drug Discovery. https://doi.org/10.1101/2024.07.25.605062

Hicks, M.T., Humphries, J., Slater, J., 2024. ChatGPT is bullshit. Ethics Inf. Technol. 26, 38. https://doi.org/10.1007/s10676-024-09775-5

Kirsher, D.Y., Chand, S., Phong, A., Nguyen, B., Szoke, B.G., Ahadi, S., 2025. Current landscape of plasma proteomics from technical innovations to biological insights and biomarker discovery. Commun. Chem. 8, 279. https://doi.org/10.1038/s42004-025-01665-1

Kusumegi, K., Yang, X., Ginsparg, P., de Vaan, M., Stuart, T., Yin, Y., 2025. Scientific production in the era of large language models. Science 390, 1240–1243. https://doi.org/10.1126/science.adw3000

Mamoshina, P., Kochetov, K., Putin, E., Aliper, A., Zhavoronkov, A., 2019. Testing for batch effect through age predictors. https://doi.org/10.1101/531863

Mamoshina, P., Kochetov, K., Putin, E., Cortese, F., Aliper, A., Lee, W.-S., Ahn, S.-M., Uhn, L., Skjodt, N., Kovalchuk, O., Scheibye-Knudsen, M., Zhavoronkov, A., 2018a. Population Specific Biomarkers of Human Aging: A Big Data Study Using South Korean, Canadian, and Eastern European Patient Populations. J. Gerontol. Ser. A 73, 1482–1490. https://doi.org/10.1093/gerona/gly005

Mamoshina, P., Volosnikova, M., Ozerov, I.V., Putin, E., Skibina, E., Cortese, F., Zhavoronkov, A., 2018b. Machine Learning on Human Muscle Transcriptomic Data for Biomarker Discovery and Tissue-Specific Drug Target Identification. Front. Genet. 9, 242. https://doi.org/10.3389/fgene.2018.00242

Messeri, L., Crockett, M.J., 2024. Artificial intelligence and illusions of understanding in scientific research. Nature 627, 49–58. https://doi.org/10.1038/s41586-024-07146-0

Nguyen, V.K., Middleton, L.Y.M., Huang, L., Zhao, N., Verly, E., Kvasnicka, J., Sagers, L., Patel, C.J., Colacino, J., Jolliet, O., 2023. Harmonized US National Health and Nutrition Examination Survey 1988-2018 for high throughput exposome-health discovery. medRxiv 2023.02.06.23284573. https://doi.org/10.1101/2023.02.06.23284573

Putin, E., Mamoshina, P., Aliper, A., Korzinkin, M., Moskalev, A., Kolosov, A., Ostrovskiy, A., Cantor, C., Vijg, J., Zhavoronkov, A., 2016. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging 8, 1021–33. https://doi.org/10.18632/aging.100968

Rafikova, E., Nemirovich-Danchenko, N., Ogmen, A., Parfenenkova, A., Velikanova, A., Tikhonov, S., Peshkin, L., Rafikov, K., Spiridonova, O., Belova, Y., Glinin, T., Egorova, A., Batin, M., 2024. Open Genes-a new comprehensive database of human genes associated with aging and longevity. Nucleic Acids Res. 52, D950–D962. https://doi.org/10.1093/nar/gkad712

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., Barsoum, E., 2025. Agent Laboratory: Using LLM Agents as Research Assistants. https://doi.org/10.48550/arXiv.2501.04227

Shayakhmetov, R., Kuznetsov, M., Zhebrak, A., Kadurin, A., Nikolenko, S., Aliper, A., Polykovskiy, D., 2020. Molecular Generation for Desired Transcriptome Changes With Adversarial Autoencoders. Front. Pharmacol. 11. https://doi.org/10.3389/fphar.2020.00269

Sidorenko, D., Pushkov, S., Sakip, A., Leung, G.H.D., Lok, S.W.Y., Urban, A., Zagirova, D., Veviorskiy, A., Tihonova, N., Kalashnikov, A., Kozlova, E., Naumov, V., Pun, F.W., Aliper, A., Ren, F., Zhavoronkov, A., 2024. Precious2GPT: the combination of multiomics pretrained transformer and conditional diffusion for artificial multi-omics multi-species multi-tissue sample generation. Npj Aging 10, 37. https://doi.org/10.1038/s41514-024-00163-3

Singh, B., 2025. Epistemic Destabilization: AI-Driven Knowledge Generation and the Collapse of Validation Systems. Proc. AAAIACM Conf. AI Ethics Soc. 8, 2387–2398. https://doi.org/10.1609/aies.v8i3.36724

Tim Pyrkov, 2022. pynhanes: Python parser and scraper for NHANES accelerometry and questionnaire.

Urban, A., Sidorenko, D., Zagirova, D., Kozlova, E., Kalashnikov, A., Pushkov, S., Naumov, V., Sarkisova, V., Leung, G.H.D., Leung, H.W., Pun, F.W., Ozerov, I.V., Aliper, A., Ren, F., Zhavoronkov, A., 2023. Precious1GPT: multimodal transformer-based transfer learning for aging clock development and feature importance analysis for aging and age-related disease target discovery. Aging 15, 4649–4666. https://doi.org/10.18632/aging.204788

Zhavoronkov, A., Kochetov, K., Diamandis, P., Mitina, M., 2020. PsychoAge and SubjAge: development of deep markers of psychological and subjective age using artificial intelligence. Aging 12, 23548–23577. https://doi.org/10.18632/aging.202344

Zhavoronkov, A., Mamoshina, P., 2019. Deep Aging Clocks: The Emergence of AI-Based Biomarkers of Aging and Longevity. Trends Pharmacol. Sci. 40, 546–549. https://doi.org/10.1016/j.tips.2019.05.004