Model Size vs. Average Phenotype Performance

Pearson r = 0.736

Model Parameters vs. Balanced Accuracy (All Phenotypes Average)

Data points labeled: gpt-4, deepseek-chat, deepseek-r1-0324, wizardlm-2-8x22b, gpt-oss-120b, llama-3-sonar-large-32k-instruct, llama-3-70b-instruct:nitro, gpt-3.5-turbo-0125, gpt-oss-20b, gemma-7b-it, mixtral-8x7b-instruct:nitro, phi-3-mini-128k-instruct

Organization
- DeepSeek
- Google
- Meta AI
- Microsoft
- Mistral AI
- OpenAI
- Perplexity
- Tsinghua University
- Linear fit