INTELLIGENCE BANDWIDTH

Haifeng Jin
Independent Researcher

ABSTRACT

Traditional benchmarks for artificial intelligence (AI) have predominantly focused on the quality and accuracy of outputs, largely overlooking the critical dimension of processing speed. This omission is particularly problematic, as many real-world AI applications, from autonomous driving to customer service, are time-sensitive. Existing speed metrics such as Tokens Per Second (TPS) are insufficient, as they are token-centric regardless of usefulness and ill-suited for a multi-modal future. This paper introduces "Intelligence Bandwidth" as a new metric to measure the processing speed of AI, defined as the amount of useful information an AI can produce per unit of time. Several methods for its approximation are proposed, with a focus on measuring raw output bits per second for its simplicity and modality-agnostic nature. By analyzing historically significant generative AI models, a clear trend of exponential growth is observed. From this data, "Jin's law" is formulated, positing that the intelligence bandwidth of the best publicly available AI model doubles approximately every year. This law provides a predictive framework for the evolution of human-AI interaction, forecasting the near-term integration of real-time image generation into text-based conversations and the advent of real-time video interaction within the next three years.

1 INTRODUCTION

Processing speed has long been recognized as an important metric in human cognitive ability [38] and intelligence scale measurement [9, 19, 8]. It reflects the fundamental capacity to perform daily tasks as a human, including reading comprehension, communication, and driving. Both the quality of work and the time required are essential to its evaluation.

To measure the intelligence scale of artificial intelligence (AI), various benchmark datasets have been designed, analogous to tests for humans. For example, the MMLU dataset [13] consists of multiple-choice questions to test multi-language understanding, the GPQA dataset [27] consists of graduate-level questions on a variety of subjects, and the MATH dataset [14] comprises competition math problems.

However, when experiments are conducted to test AI on these datasets, unlike tests for humans, only the quality of work is emphasized, while the time aspect of the test is typically ignored. Such an approach does not satisfy the requirements of AI in real-world applications.

From a pragmatic perspective, many tasks are time-sensitive, such as autonomous driving and customer service. In these applications, the output of intelligence is required within a limited time window. Even in the early days of deep learning, there was an implicit assumption of a time limit. When AlphaGo [30] defeated the best human Go player in 2016, it adhered to the same rules, including time constraints, as the human player.

1.1 IMPLICATIONS OF TEST-TIME SCALING

With the advent of the test-time scaling law [31], the scores achieved on these AI benchmark datasets alone can no longer fully measure intelligence.

According to the test-time scaling law, a small model can solve harder problems by spending more tokens "thinking," while a larger model can solve the same problem with fewer tokens. By analogy to a human IQ test, it is as if one person spent more time and used a long sheet of scratch paper to finish the test, while another person used very little time and no scratch paper at all. If both achieve

the same number of correct answers, it is more rational to conclude that the person who completed the test faster possesses higher intelligence, likely employing a more efficient method.

That smaller and larger models can answer the same question correctly does not imply they have the same intelligence level; rather, they approached the problem differently. Therefore, without any constraint on the output—whether in terms of the number of tokens or output speed—it is impossible to measure the actual intelligence level of AI solely based on correctness.

From a user's perspective, estimating the time required for AI to complete tasks has become increasingly challenging. Prior to the advent of test-time scaling, models typically used a similar number of tokens to solve each problem. With test-time scaling, models are incentivized to generate more tokens for certain problems, leading to greater variability. Users now have access to metrics such as tokens per second for the APIs they utilize, but lack information about the total number of tokens needed for a given task. As a result, they can no longer reliably estimate the time required for task completion.

1.2 WHY SPEED WAS IGNORED

If processing speed is so important to intelligence, why was it historically ignored? The reasons are twofold. First, AI was not sufficiently advanced. Early systems could not pass the Turing test [36], solve math or coding problems, or perform in-depth reading or writing. Given that AI was in a primitive stage, the focus was on enabling AI to perform new tasks rather than on the speed of task completion. Second, AI applications were all task-specific. The AI for X-ray image processing [6] was entirely different from the AI used for recommendation systems [12]. The applicability of AI was measured case by case by application builders, without the need for a unified processing speed metric.

However, with the advent of large language models (LLMs) [4], AI has become substantially more capable and generalizable, with applications spanning from medical diagnosis [37] to recommendation systems [39]. Consequently, it is now both rational and timely to establish a unified metric for AI processing speed.

1.3 A MENTAL SHIFT FOR EVALUATING AI

Additionally, there has been a significant shift in how application builders interface with models as summarized in Table 1. Initially, developers focused on training application-specific deep learning models, beginning with the advent of AlexNet [17]. With the emergence of ChatGPT [4], the paradigm shifted toward pretraining large language models (LLMs). The public release of Llama [35] enabled post-training and fine-tuning of open models. As the capabilities of open models advanced, such as the Qwen series [3], it became increasingly feasible to serve an open-weight model "as is" without additional fine-tuning. More recently, the introduction of DeepSeek [11], which substantially reduced token costs, has made it more cost-effective to utilize hosted APIs from major AI service providers rather than maintaining proprietary infrastructure.

Table 1: Paradigm shifts of AI usages

Paradigms	**Defining Moments**
Deep Learning	AlexNet
Pretraining LLMs	ChatGPT
Open Model Fine-tuning	Llama
Open Model Serving	Qwen
Hosted APIs	DeepSeek

When evaluating the intelligence of AI, a shift is required from evaluating static models, pure mathematical constructs defined by neural architectures and parameters, to evaluating hosted AI services or APIs. It is necessary to assess their processing speed for applicability to time-sensitive tasks.

2 RELATED WORK

There are established metrics to evaluate the speed of LLMs, such as time to first token (TTFT) and tokens per second (TPS) [40]. These are suitable metrics for serving language models, but there are two major limitations when considering them as general metrics for AI processing speed.

First, unlike the time-limit for an IQ test, these metrics focus on the number of tokens rather than the actual useful information produced by intelligence. For example, a simple program without any intelligence can output random tokens at a very high speed.

Second, they are not designed for a multi-modal future of AI. Although the term multi-media may seem outdated, it is highly relevant to contemporary AI development. LLMs have largely pushed human-computer interaction back to the pre-multi-media era. On modern social media, users read articles with images and watch long, short, or livestream videos. In contrast, current AI interactions are predominantly text-based, reminiscent of the early 1990s internet.

With the emergence of multi-modal AI technologies, such as image generation [16, 10, 28] and video generation [23, 15, 24], it is anticipated that the future of human-AI interaction [1] and even AI-AI interaction [33] will be multi-modal.

What is needed is a metric that quantifies the rate at which useful information is produced, rather than merely counting tokens, and that remains applicable across diverse modalities to ensure future relevance.

3 INTELLIGENCE BANDWIDTH

This paper introduces the concept of intelligence bandwidth as a metric for the processing speed of AI.

Definition 1. *Intelligence bandwidth* is a measurement indicating the maximum amount of useful information that a served artificial intelligence model can produce in a given amount of time, formally expressed as:

$$B = \frac{I}{t}$$

where B is the intelligence bandwidth, I is the amount of useful information, and t is the total time spent.

It is important to note that intelligence bandwidth primarily measures the output speed of AI, rather than input, for two reasons. First, the main impact of intelligence bandwidth is in human-AI interaction, which will be discussed further in the next section. Second, AI's processing speed is mainly bounded by output, not input. Thus, it is more meaningful to track AI progress with a metric that measures its bottleneck.

4 APPROXIMATED MEASUREMENTS

Despite the formal definition of intelligence bandwidth, certain ambiguities remain in practical measurement. First, "useful information" is not an unambiguous term. Defining the usefulness of an AI model remains a challenging task. Second, time is also subject to multiple interpretations in computer science.

Regarding time measurement, the two popular choices are CPU time and wall time. Wall time is preferred, as intelligence bandwidth is primarily relevant for human-AI interactions and applications. End-to-end latency is more informative than technical details such as CPU time.

Regarding useful information, the evaluation is based solely on the interaction between the user and AI. The downstream use of the output is not considered. For example, if a user asks "what is the square of 3?" and uses the answer in a math competition to win a '$10k' bonus, the usefulness is measured by the quality of the AI's answer, not the subsequent economic gain.

To address the ambiguity in "useful information" in Definition 1, three approximate methods for calculating intelligence bandwidth are proposed. These methods are all approximations, as there is no universally agreed-upon way to measure useful information.

4.1 METHOD 1: BENCHMARK SCORE DIVIDED BY TIME

The first method is to divide the score achieved by AI on benchmark datasets by the time required to produce the answer. In this approximation, usefulness is measured by the score on benchmarking datasets, formally expressed as follows.

Let $S = \{s_1, s_2, \ldots, s_n\}$ be a set of scores from n individual AI benchmarks normalized to the same range, and $W = \{w_1, w_2, \ldots, w_n\}$ be a corresponding set of weights, where w_i represents the relative importance of each benchmark. The total time expended across all assessments is denoted by t. Intelligence bandwidth B can be approximated as:

$$B = \frac{\sum_{i=1}^{n} w_i s_i}{\sum_{i=1}^{n} w_i \cdot t}$$

For an AI application developer with a time-bounded task, the tokens/second metric provides limited insight, as the number of tokens required for the task is unknown. However, this approximation of intelligence bandwidth offers a more informative measure of how much useful work can be accomplished per second.

4.2 METHOD 2: THE INFORMATION THEORY APPROACH

The second method is based on information theory. This approach explores whether the foundational framework of information theory [29] can be used to measure the amount of information output by an AI model.

Intelligence bandwidth B can be approximated as:

$$B = \frac{I'(X)}{t}$$

where X is the output of AI, $I'(X)$ is the information content of X. In the context of LLMs, $I'(X)$ of a sentence $X = (x_1, x_2, \ldots, x_n)$ is given by:

$$I'(X) = -\sum_{i=1}^{n} \log_2 P(x_i | x_1, x_2, \ldots, x_{i-1})$$

where x_i is the i-th word in the output sentence. $P(x_i | x_1, x_2, \ldots, x_{i-1})$ is the conditional probability of the i-th word occurring, given all preceding words. For other modalities, such as audio, image, and video, appropriate measures of information content should be used.

There are two limitations to this approximation:

1. It requires access to the probability output by the large language model. The value of P is contained in the output probability vector.
2. The amount of information does not necessarily equate to the amount of useful information. For example, a simple GPT-2 [26] model can output text rapidly but with limited usefulness.

Therefore, this method is less widely applicable or accurate than the first method.

4.3 METHOD 3: RAW OUTPUT BITS

The third approximate method is to measure the number of bits in the raw outputs of the AI model. With this method, intelligence bandwidth is measured in bits per second.

The advantage of this method is its simplicity. The number of bits in any text, image, or video output by an AI model can be easily computed without using any benchmark dataset or accessing the output probability vector.

Another advantage is the absence of ambiguity, such as the selection of benchmark datasets in the first method or the computation of probabilities in the second method.

The primary limitation of this approach is that it diverges from measuring the actual usefulness of the information output by the AI model. However, if the models measured are constrained to the best available in the market, the useful information per bit should be similar among them. Therefore, this approximation is valid in a constrained environment.

5 Impact on human-AI interaction

Effective metrics can help reveal new laws to predict the future, analogous to feature width and Moore's law [20], or internet bandwidth and Nielsen's law [22]. Examining the history of the internet through the lens of network bandwidth, as bandwidth increased, richer content formats emerged [7], shifting from text-based websites like early Twitter [21] to video-based platforms like YouTube [32]. With a simple metric such as network bandwidth, it was possible to predict the popularity of certain applications. Similarly, a robust metric for AI processing speed can facilitate the discovery of macro-level trends in AI development.

Intelligence bandwidth, as a metric, tracks the progress of human-AI interaction. Current human-AI interactions are still predominantly text-based. This is possible because the output speed of AI has exceeded human reading speed, which is approximately 238 words per minute [5]. This is enabled by state-of-the-art serving technologies, which can generate 14,000 words per minute. Similarly, speech generation speed is far beyond human listening speed [18].

To clarify the prerequisites for enabling real-time human-AI interaction within a given modality, it is helpful to examine the specific conditions that must be satisfied. For self-paced media formats such as text and images, individuals typically consume content at their maximum perceptual speed. Once the intelligence bandwidth exceeds the human perceptual threshold, real-time interaction in that modality becomes feasible. In contrast, for fixed-speed media formats such as audio and video, users generally adhere to the inherent playback speed. Thus, as long as the AI generation speed surpasses the fixed playback rate of these media formats, real-time interaction in those modalities is achievable.

Research in multi-modal AI continues to address the bottlenecks of other modalities. With the increase of the intelligence bandwidth of image generation, visual illustrations will become integrated into AI responses. AI may also be able to perform visual reasoning akin to humans on a whiteboard, and iteratively refine graphical designs as designers do on paper. As the intelligence bandwidth of leading AI models continues to increase, it is expected that video illustrations and real-time generated environment interactions, such as those demonstrated by Genie 3 [25], will become feasible. In a speculative future, AI could generate entire worlds that users can interact with in real time—an idealistic scenario enabled by the cognitive capabilities of AI.

Achieving such increases in intelligence bandwidth requires not only advances in AI models as static collections of neural architectures and parameters, but also significant improvements in AI hardware and machine learning systems. In this vision, hardware, software, and models are no longer orthogonal, but are deeply integrated to enable superintelligence.

6 Experiments

In this section, all historically significant generative AI models are measured for their intelligence bandwidth and plotted in a single figure. The models covered include large language models, image generators, and video generators. The raw output bits method is used due to its ease of measurement and minimal ambiguity. Most of the data presented in this section is collected from Artificial Analysis [2].

The experimental results are shown in Figure 1, where the X-axis is the release date of the models, and the Y-axis is the intelligence bandwidth of the models measured in kilobytes per second. The modality of the models is indicated by different colors as shown in the legend.

Key observations from the experimental results include:

1. Most language models are between 0 KB/s and 3 KB/s.
2. Image generators exhibit an exponential growth rate.

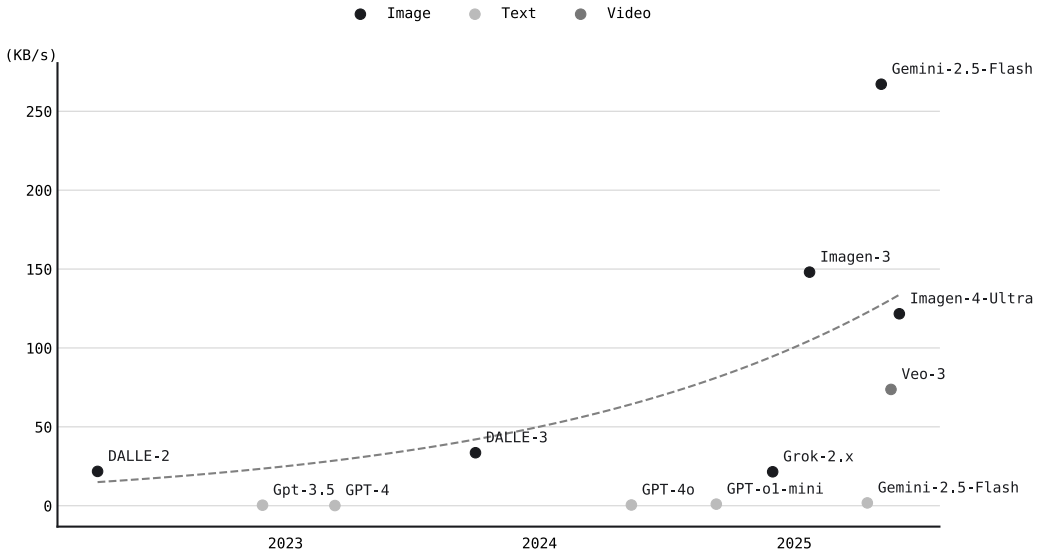

Figure 1: Intelligence bandwidth (KB/s) over time.

3. The video generator, Veo3 [15], currently exhibits an even lower intelligence bandwidth than the state-of-the-art image generators. This is primarily attributable to less mature serving technologies for video models compared to those for large language models and image generators. As serving efficiency for video generators improves, substantial growth in their intelligence bandwidth is anticipated in the near future.
4. The Gemini 2.5 Flash [34] image generator is an outlier, primarily because it is optimized for low latency and usability rather than best quality and fidelity.

7 JIN'S LAW

As we mentioned above, a robust metric for AI processing speed can facilitate the discovery of macro-level trends in AI development. We now assess the validity of intelligence bandwidth as such a metric by examining whether it supports the formulation of a predictive law for future AI growth.

The dotted curve in Figure 1 represents the estimated growth of intelligence bandwidth. The prediction of the growth rate is based primarily on Imagen 4, the state-of-the-art high-quality image generator, rather than a model balanced between speed and quality, such as Gemini 2.5 Flash [34]. The growth rate of intelligence bandwidth is summarized in a simple law named after the author's surname, presented as follows.

Jin's law: The intelligence bandwidth (KB/s) of the best hosted AI model available to the public doubles every year.

The formal definition of this law is as follows. Let $B(t)$ be the intelligence bandwidth of the best hosted AI model available to the public at time t and B_0 be the initial intelligence bandwidth at time t_0. The law can be expressed as:

$$B(t) = B_0 \cdot 2^{(t-t_0)/T}$$

where T is the doubling period. The best estimate is $T = 1$, the unit of which is year.

In Jin's law, intelligence bandwidth is defined by the modality exhibiting the highest KB/s measurement. Currently, image generators are at the forefront of this growth. As advancements in models and serving technologies for image generation reach a plateau, it is anticipated that video generators will become the primary drivers of further increases in intelligence bandwidth.

Based on Jin's law, two predictions about human-AI interaction in the near future are made as an example of how to use the law to predict the exponential growth of AI in the future:

1. **Images will soon be used in AI interactions.** The latency of the Gemini 2.5 Flash [34] image generator is lower than that of large language model responses. Consequently, large language models may soon incorporate images to provide enhanced illustrations in their outputs. Currently, it takes only 4.6 seconds to generate an image. If this speed doubles within a year, it is likely that applications will emerge in which images become a primary mode of interaction and illustration.
2. **Real-time video interaction will be widely available in three years.** Given the intelligence bandwidth of models in 2025, it is currently possible to generate 8 seconds of video in 50 to 60 seconds. If generation speed increases by a factor of 7 to 8, real-time video generation will become feasible. Achieving an 8-fold increase corresponds to approximately $\log_2(8) = 3$ years. While some uncertainty remains in this prediction, video generators are presently below the projected growth curve and possess significant potential for accelerated improvement as serving technologies advance.

There are many other implications that can be derived from Jin's law. It is hoped that this law will guide AI application developers in identifying optimal time windows to bring products to market, and policymakers in enacting regulations at appropriate times to maximize development while minimizing harm.

8 LIMITATIONS

There are several limitations to this work. First, the measurement of usefulness remains an open challenge. This paper proposes three straightforward approaches to approximate usefulness in AI outputs; however, more rigorous and comprehensive methods are needed. Second, the accuracy of the estimated doubling period T in Jin's law is uncertain. As the field of AI is still in its early stages, the current estimation is based on a limited number of available data points. Third, the exponential growth described by Jin's law represents an idealized scenario. In practice, growth may be constrained by factors such as energy supply limitations or economic pressures, particularly if the AI sector experiences a market correction.

9 CONCLUSIONS

The paper introduces the concept of "Intelligence Bandwidth," a metric that measures the processing speed of served AI models by quantifying the rate at which they produce useful information. The paper observes an exponential growth trend in this metric across various modalities and formulates "Jin's law," which states that the intelligence bandwidth of the best publicly available AI doubles annually. This law provides a predictive framework, forecasting the imminent integration of real-time images into human-AI text interactions and the widespread availability of real-time video generation within three years. The paper advocates for a new evaluation paradigm that focuses on dynamic, served AI systems, emphasizing the crucial interplay between hardware, software, and models.

Ultimately, the most important takeaway is that when evaluating AI for roles traditionally performed by humans, the evaluation criteria should mirror those used for humans. The distinction between assessing human and AI performance for a given task will become increasingly blurred. Innovative approaches are required to integrate processing speed into comprehensive intelligence assessments for AI.

ACKNOWLEDGMENTS

Completing this work required stepping outside established research areas, a challenge that would not have been possible without the encouragement and support of many individuals. Sincere gratitude is extended to everyone who provided guidance and encouragement during this process. Your belief in this project and in the ability to see it through was invaluable.

REFERENCES

[1] Saleema Amershi et al. Guidelines for human-ai interaction. In *CHI*, 2019.

[2] Artificial Analysis, Inc. AI Model and API Providers Analysis. https://artificialanalysis.ai/, 2024.

[3] Jinze Bai et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[4] Tom Brown et al. Language models are few-shot learners. In *NeurIPS*, 2020.

[5] Marc Brysbaert. How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 2019.

[6] Erdi Çallı et al. Deep learning for chest X-ray analysis: A survey. *Medical Image Analysis*, 2021.

[7] Kerry G Coffman et al. Internet growth: Is there a "Moore's Law" for data traffic? *Handbook of massive data sets*, 2002.

[8] Dawn P Flanagan et al. *Essentials of WISC-V assessment*. John Wiley & Sons, 2017.

[9] Astrid F Fry et al. Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, 1996.

[10] Ian J Goodfellow et al. Generative adversarial nets. *NeurIPS*, 2014.

[11] Daya Guo et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[12] Xiangnan He et al. Neural collaborative filtering. In *The Web Conference*, 2017.

[13] Dan Hendrycks et al. Measuring massive multitask language understanding. In *ICLR*, 2021.

[14] Dan Hendrycks et al. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS*, 2021.

[15] Tom Hume et al. Meet Flow: AI-powered filmmaking with Veo 3. https://blog.google/technology/ai/google-flow-veo-ai-filmmaking-tool/, 2025.

[16] Diederik P Kingma et al. Auto-encoding variational Bayes. In *ICLR*, 2014.

[17] Alex Krizhevsky et al. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012.

[18] Victor Kuperman et al. A lingering question addressed: Reading rate and most efficient listening rate are highly similar. *Journal of Experimental Psychology: Human Perception and Performance*, 2021.

[19] Elizabeth O Lichtenberger et al. *Essentials of WAIS-IV assessment*. John Wiley & Sons, 2012.

[20] Gordon E Moore et al. Cramming more components onto integrated circuits, 1965.

[21] Dhiraj Murthy. *Twitter*. Polity Press Cambridge, 2018.

[22] Jakob Nielsen. Nielsen's law of internet bandwidth. http://www.useit.com/alertbox/980405.html, 1998.

[23] Oord, Aäron van den et al. State-of-the-art video and image generation with Veo 2 and Imagen 3. https://blog.google/technology/google-labs/video-image-generation-update-december-2024/, 2024.

[24] OpenAI. Sora: System card. https://openai.com/index/sora-system-card/, 2024.

[25] Jack Parker-Holder et al. Genie 3: A new frontier for world models. https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/, 2025.

[26] Alec Radford et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[27] David Rein et al. GPQA: A graduate-level Google-Proof Q&A benchmark. In *COLM*, 2024.

[28] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[29] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 1948.

[30] David Silver et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 2016.

[31] Charlie Snell et al. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *ICLR*, 2025.

[32] Pelle Snickars and Patrick Vonderau. *The YouTube reader*. Kungliga biblioteket, 2009.

[33] Rao Surapaneni et al. Announcing the Agent2Agent Protocol (A2A). `https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/`, 2025.

[34] Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[35] Hugo Touvron et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[36] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[37] Ehsan Ullah et al. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic Pathology*, 2024.

[38] Richard W Woodcock et al. *Woodcock-Johnson tests of cognitive ability*. DLM Teaching Resources, 1989.

[39] Xuansheng Wu et al. Could small language models serve as recommenders? Towards data-centric cold-start recommendation. In *The Web Conference*, 2024.

[40] Yinmin Zhong et al. DistServe: disaggregating prefill and decoding for goodput-optimized large language model serving. In *OSDI*, 2024.