Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL33655

Task-Specific Efficacy of Contemporary Large Language Models: A Comparative Survey of ChatGPT, DeepSeek, Gemini, Qwen, and LLaMA

Weijie Shangguan

College of Information Engineering, Xi'an Eurasia University, Xi'an, China
sgwj0825@outlook.com

Abstract. The past few years have brought a flood of new large language models. Each one arrives with its own design philosophy and strengths, which makes it tough for working professionals to figure out which tool actually fits their daily tasks. Standard test scores do not always point to the right answer. This paper takes a close look at five widely used systems. They are ChatGPT, DeepSeek-V3, Gemini 2.5, Qwen3, and LLaMA. The analysis draws on what the developers themselves have published and what outside researchers have found in controlled experiments. One thing becomes clear right away. These tools have divided up the work in interesting ways. DeepSeek-V3 tackles science computing and coding tasks more effectively than ChatGPT, and it runs at a much lower cost. Gemini 2.5 proves its worth when the job demands handling very long documents or mixing together pictures, sound, and text. Qwen3 pulls ahead in translation work across many languages and in building the parts of software that users see and touch. ChatGPT holds onto its spot as the favorite for spinning stories and cooking up new ideas. LLaMA has grown into a home base for teams that want to craft their own custom tools. The takeaway from the research is straightforward. Choose the tool based on the task sitting in front of people, not on a number from some public leaderboard.

Keywords: Large language models, Model benchmarking, Task-model matching, Mixture-of-Experts, Domain-specific AI

1. Introduction

The arrival of ChatGPT near the end of 2022 shook up how a lot of people think about what machines can do with words. A wave of new systems followed right on its heels. Meta put LLaMA into the open-source world early in 2023. That single move let thousands of researchers around the globe start building things together [1]. Google rolled out Gemini 1.0 later that same year. The team behind it set out to build something that could chew on text, images, and sound all at the same time, not as a feature tacked on after the fact [2]. They then pushed its memory out to hold two million tokens with Gemini 1.5 [3]. After that came deeper reasoning skills and the power to carry out multi-step jobs on its own in Gemini 2.5 [4]. DeepSeek headed down a different road at the end of 2024. The V3 system leans on a setup called Mixture-of-Experts. Out of 671 billion total parameters, only

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL33655

37 billion wake up for any given query. That design choice slashes how much each run costs [5]. Alibaba stepped into the ring in 2025 with Qwen3, a system built to juggle 119 languages and chew through long documents [6].

Having so many options sounds great, but it makes picking the right one a real headache. Someone who writes code from morning to night needs a very different assistant than someone who translates academic papers or stitches together video projects. The leaderboards everyone checks, like MMLU or Chatbot Arena, hand out a number. That number does not whisper much about how the tool will handle an actual workload on a rainy Wednesday. What people need is a way to line up tasks with tools based on hard evidence. That is the hole this paper tries to plug.

A handful of research groups have already laid some of the bricks for this kind of work. One team pulled together results from a wide range of studies. They showed that open models now go toe-to-toe with the ones locked behind paywalls [7]. Another group ran science computing tests and watched DeepSeek-V3 beat ChatGPT at churning out physics simulation code. The cost difference was just as striking, running somewhere between a tenth and a twentieth of what ChatGPT charged per query [8]. A third team threw seven systems at academic writing jobs. Qwen3 came out on top for Chinese-to-English translation. DeepSeek-V3 stumbled less often on LaTeX formatting. ChatGPT wrote the most natural-sounding original passages [9]. Each of these findings adds a piece to the puzzle. No one has yet snapped all the pieces together into a picture that a busy professional could actually lean on.

This paper makes an attempt at that. The analysis looked hard at the five systems named above, tying what the studies say about their strengths to the kinds of jobs people grind through every day. For this review, the scholarly literature available through Google Scholar was searched. The search covered materials published between 2023 and 2025. Technical reports straight from the model builders and independent benchmarking work that stacks up at least two of the five systems against each other were the focus.

Section 2 walks through each system and what hums under its hood. Section 3 sets them side by side on real flavors of work. Section 4 digs into what this all means for picking the right tool and where the field might drift next. Section 5 wraps things up with the core lessons.

2. Overview of relevant technologies

2.1. ChatGPT

ChatGPT made its public bow in November 2022. OpenAI propped it up on the GPT-3.5 frame and trained it to cough up better answers by having human graders score what came out [1]. GPT-4 showed up four months later. It could size up images about as smoothly as it read text. Underneath sits what the engineers call a dense Transformer. Every single parameter in the thing fires off for every single job. Its reading span has stretched from an early cap of four thousand tokens to a hundred twenty-eight thousand today. Folks grab for it because it keeps a conversation rolling, cooks up creative stuff across all sorts of formats, and can walk someone who has never coded through a tangled chunk of logic in everyday words. The downsides are worth noting. Calling it through an API drains the wallet faster than newer rivals, and tests have caught it fumbling on math that demands dead-on answers [9].

DOI: 10.54254/2755-2721/2026.GL33655

2.2. DeepSeek-V3

DeepSeek-V3 landed near the tail end of 2024. The big story swirling around it is thrift. Earlier spins on DeepSeek had already shown that a Mixture-of-Experts blueprint could work at real scale. V3 folded in something called Multi-Token Prediction and FP8 training, a method that shrinks the memory bill during the training run [6]. A quick look at the numbers tells the tale. The whole setup contains 671 billion parameters, but only 37 billion of them stir to life for any one query. A bit called Multi-head Latent Attention crunches down the memory keys so the words pour out faster. Outside testing has found it punches hard at math, at writing code, and at coughing up tidy outputs like SQL queries and LaTeX pages. The main headache for anyone itching to run it on their own hardware is the seven hundred gigabytes of video memory the full beast demands.

2.3. Gemini 2.5

Google kicked off its Gemini line in December 2023 with a clear mission. The goal was to craft something that understands text, pictures, noise, and moving images from the ground up, not as a patch glued on later [3]. By February 2024, Gemini 1.5 could cram two million tokens into its working memory [4]. The 2025 edition stacked advanced reasoning and the knack for steering through multi-step tasks without a human nudging every single move [5]. Benchmarks back up the hype. It can find a needle in a haystack across that whole two-million-token sweep with perfect marks. It also grinds through long documents, chomps on whole codebases, and sifts through mountains of log files without breaking a sweat. The catch is that people are tied to Google's cloud, and the meter runs up faster than with most other options.

2.4. Qwen3

Alibaba set Qwen3 on the table in 2025, picking up a thread of work that goes back to 2023 [7]. Like DeepSeek, it rides on Mixture-of-Experts, gulps down a million tokens at a go, and blankets 119 languages. The folks who designed it poured real sweat into balancing raw size against running costs by training on a sprawling stew of languages. Independent testing shows it topping models of similar heft at Chinese-English academic translation. That holds whether people check automated scores or ask human judges [10]. The team also honed it to help developers crank out front-end interface pieces. It still runs a step behind DeepSeek-V3 on science and tech benchmarks given only in English.

2.5. LLaMa

Meta unleashed the first LLaMA brood in February 2023. The sizes ran from seven billion up to sixty-five billion parameters. These models quickly turned into the foundation that the whole open-source world started building on [2]. The line branched out through LLaMA 2 and LLaMA 3, giving rise to thousands of specialized offshoots. Alpaca, Vicuna, Code Llama, and many more popped up, each tuned for a different corner of medicine, law, or finance. On the inside, LLaMA sticks to the familiar dense Transformer mold with a few nice tweaks like SwiGLU and rotary position encoding. Its real punch comes not from some flashy trick buried in the code. It comes from the enormous swarm of developers that has gathered around it. When a bank or a hospital wants to build a private AI tool on its own store of data, LLaMA is almost always where they dig the first hole. One thing to keep in mind is that the base version usually needs extra training on specific tasks before it can run with the big dogs in any narrow field.

DOI: 10.54254/2755-2721/2026.GL33655

3. The differences among large models and their advantageous application fields

3.1. Technical research and development: code, data, and long-document processing

On science programming and algorithm puzzles, DeepSeek-V3 has yanked ahead of the competition. Told to spit out working code for physics-informed neural nets, it nails the job more often than ChatGPT. At the same time it chews through API resources at a rate that runs ten to twenty times lower [9]. On standard coding tests like HumanEval, both systems land at roughly the same accuracy numbers. Yet DeepSeek-V3 fires back its answers quicker and at a lighter cost. Gemini 2.5 does not bring anything extra to these code-heavy jobs. The data puts DeepSeek-V3 at the front here, with both systems outrunning Gemini 2.5.

The ranking does a complete somersault when the job shifts to wrapping people's head around a whole software project or reshaping code that snakes across dozens of files. For this kind of grind, Gemini 2.5 sits in a class all by itself. That two-million-token gullet lets it swallow a mid-sized codebase in one go. Researchers are talking hundreds of thousands of lines flung across scores of files, with dependency tracing that respects every boundary [5]. Neither DeepSeek-V3 nor ChatGPT can pull off that move. Both smash into a wall around 128K tokens and must hack the code into chunks. Carving things up like that inevitably snaps the threads tying far-flung parts of the program together, hiding the big-picture design from view. Gemini 2.5 therefore owns this corner of engineering work without any real challenger.

For database slogging and data pipeline scripting, DeepSeek-V3 nudges ahead. The SQL it writes hugs closer to proper syntax, and the tuning tips it tosses out hit the mark more often than what ChatGPT serves up. Hand Gemini 2.5 a complete data dictionary and schema map, and it can burp out an entire ETL sequence in one shot. Weighing everything on the scale, DeepSeek-V3 keeps a slim but real edge over ChatGPT for the daily grind of data engineering.

The tables spin around once more when the job centers on explaining code and writing up technical docs. ChatGPT pumps out annotations that read smoother, unpack ideas more clearly, and score better across the board on readability [1]. DeepSeek-V3 goes for clipped, all-business descriptions that assume people already know a fair chunk of the backstory. When a team needs to haul new folks up the learning curve fast, ChatGPT is the smarter pick.

Stepping back from the task-by-task weeds, a crisp picture of who does what across the technical landscape snaps into view. DeepSeek-V3 has edged past ChatGPT in science coding, query building, and tidy document formatting, all while keeping a meaningful cost advantage in its pocket. Gemini 2.5 holds the keys to the kingdom when it comes to grasping whole repositories and sifting through monster log files. There is simply no real stand-in for it today. Qwen3 has scratched out a spot in front-end work through sharp, targeted fine-tuning. ChatGPT still flies the flag for code documentation quality, though the lead it once held in raw code generation has pretty much melted away.

3.2. Content creation and knowledge work: composition, translation, and cross-media understanding

In the translation ring, Qwen3 sprints out ahead of the field. Stacked up on Chinese-English academic translation, it racks up BLEU numbers and human quality scores that sit comfortably above what ChatGPT, Gemini, or matching LLaMA builds can scrape together [10]. Its built-in grip on 119 languages hands it both the widest net and the most natural ring of any system people sized

DOI: 10.54254/2755-2721/2026.GL33655

up [7]. For scholars and editors whose work keeps hopping across language lines, Qwen3 is the strongest hand on the table right now.

When the job calls for yanking together threads of insight from a mountain of papers, the grinding kind of synthesis that sits at the heart of a solid literature review, Gemini 2.5 again peels away from the pack. It can chew through dozens of full-text articles in a single sweep and faithfully sketch the web of ideas stitching them together. ChatGPT and DeepSeek-V3 just cannot work this corner. Their tighter intake windows force them to nibble piece by piece, a method that regularly severs cross-references and tears the flow of arguments that run from one paper into the next [5]. For scholars who need deep, wide synthesis, Gemini serves up something no other current tool can dish out.

Creative writing, stories, persuasive essays, the opening paragraphs of a paper, stays ChatGPT's turf. Study after study logs stronger showings on argumentative wallop, character heft, narrative pull, and stylistic sparkle [10]. In loose, freewheeling brainstorm sessions, it also flashes more finesse than DeepSeek-V3, which tends to lock onto one convergent answer, or Gemini 2.5, whose muscle lies more in dredging up known facts than in spinning fresh cloth. When the gig rewards inventiveness over number-crunching, ChatGPT keeps the crown.

Cross-media understanding rips open the widest gaps between these systems at the architectural level. Gemini 2.5's true multimodal guts, where cross-channel links were forged during training, not screwed on afterward, give it a heavy upper hand when gnawing on stuff that stirs together prose, math notation, and charts. The reasoning trails it yanks from these hybrid documents run far thicker and more on target than what ChatGPT can manage with its tacked-on visual piece dangling off a text-centered core [5]. DeepSeek-V3, being nothing but text, cannot lay a finger on visual content at all. The gulf yawns even wider with video. Gemini 2.5 stands as the only system in the bunch that can gulp down a video file straight and follow what unfolds over time to spit out synced captions [5]. The pecking order here leaves no wiggle room. Gemini 2.5 runs way out in front of ChatGPT, which in turn keeps a clear lead over DeepSeek-V3.

Pulling the whole content and knowledge picture into one frame, Gemini 2.5 rules the turf of literature synthesis, video sense-making, and mixed-media picking apart. Qwen3 has thumped both ChatGPT and Gemini in Chinese-English academic translation and multilingual content churn, locking in its gig as the go-to for language-diverse slogging. ChatGPT hangs onto its edge over DeepSeek-V3 in creative writing and idea spinning, its conversational polish still the mark to shoot for. DeepSeek-V3 has crept past ChatGPT in churning out long technical how-tos, holding its logical thread straighter across stretched-out instructional docs.

4. Discussion

A stubborn ditch runs between what standard test kits measure and what real professional toil actually asks for. The assessment tools folks grab for most often, with MMLU and HumanEval sitting at the top of that pile, are getting gummed up by test data leaking into training sets and scores bunching up near the ceiling. Both of those headaches make it tougher to tell one system from another. On top of that, ranking tables tend to paper over hefty swings in how a system behaves once people yank it out of tidy, curated test beds and drop it into the rough and tumble of actual working conditions.

The stuff littering today's leaderboards barely brushes the surface of what professionals truly need. Cooking up reliable scientific code, tracing how pieces depend on each other across a whole software project, turning out spot-on academic translation between languages whose grammar bones are nothing alike, none of these are the kinds of things current benchmarks catch well. What the

DOI: 10.54254/2755-2721/2026.GL33655

field desperately needs are test sets purpose-built for particular jobs, kept behind walls where training data cannot seep in, and shaped to mirror the grit of how people actually use these tools day after day.

4.1. Challenges: the evaluation gap between benchmarks and real-world tasks

The gap between leaderboard rankings and actual utility is a major hurdle. High scores on general tests do not guarantee that a model can handle the specific nuances of scientific computing or academic translation.

4.2. Deployment trade-offs and model selection recommendations

The heap of evidence piled up through this dig points to some no-nonsense rules for grabbing the right tool. For science code, algorithm grinds, and knotty database queries, DeepSeek-V3 rises to the top of the heap. It goes blow for blow with ChatGPT on accuracy across the matching job families, all while hacking per-run infrastructure costs down by a factor of ten to twenty [9]. When the gig calls for wrapping people's head around a whole repository, exhaustive digging through scattered system logs, pulling together findings from a stack of full research papers, or working through video content, Gemini 2.5 stands alone. No other rig on the market right now matches that kind of intake muscle hitched to genuine multimodal handling [5].

Front-end building tilts toward Qwen3, a system that has stacked up real gains through targeted fine-tuning aimed dead at that sort of work. Jobs that stretch across multiple languages, anything from scholarly doc translation to whipping up content for crowds that speak all sorts of tongues, also lean Qwen3's way. That leaning mirrors both its wide language net and its battle-tested translation chops [7, 10]. Creative writing, idea churn, and crisp code docs stay in ChatGPT's corner [1]. Technical how-to writing and tidy doc formatting have swung over to DeepSeek-V3, which hangs on tighter to logical thread across long stretches of explanatory text [9, 10]. Multimodal content picking-apart goes to Gemini 2.5, with ChatGPT waiting in the wings as a workable but clearly thinner fallback. When an outfit needs to keep things running inside its own fences, lock down sensitive stuff, or dig deep into a narrow industry niche, the fine-tuning path paved with LLaMA offers the smartest road ahead [2].

4.3. Future directions and application prospects

Three currents look ready to rattle how these systems get put to work in the stretch ahead. The first is the climb of agentic AI, machines that can map out, fetch, stitch together, and carry through multi-step jobs on their own steam. Gemini 2.5 already has some of this baked right in, letting it steer long chains of moves without a human tapping the wheel at each turn [5]. The second current is the boom in field-specific model variants, built on LLaMA and Qwen footings to serve tight professional corners. The third is multimodal and long-context chops soaking into every cranny of the working world. The cross-modal ease and two-million-token sweep of Gemini 2.5 should be read not as the finish tape but as an early flicker that such powers are on track to go from market calling cards to everyday expectations.

Peering ahead, ChatGPT looks set to dig into its spot as the main talk-back interface, spreading deeper into tutoring, team creative work, and spur-of-the-moment idea spinning, spaces where easy chat and slick interaction keep paying off over the long haul. DeepSeek-V3, with its mix of number-crunching punch and running thrift, seems pointed toward fat uptake in STEM classrooms, science

DOI: 10.54254/2755-2721/2026.GL33655

computing dens, and sprawling code production lines, most of all where the cost of each inference bite holds the purse strings tight. Gemini 2.5 carries the heft to shake up how pros work through research synthesis, legal doc sifting, medical image matching, and cross-media content cranking, fields that have pushed back against machines exactly because they ask people to fuse stuff that shows up in wildly different shapes. Qwen3 is on a track to turn into a must-have for multilingual scholarly back-and-forth, across-border content fitting, and the worldwide slosh of know-how, above all in language crowds that have gotten the short end from English-first building sprees. LLaMA will hang onto its own odd, long-lived corner, less a head-on fighter for end-user turf than the frame beneath a vast number of those end-user apps.

5. Conclusion

This paper has trudged through five of the most buzzed-about large language models and sized up how they measure against one another on the kinds of tasks people actually face. Each system drags something of its own to the bench. DeepSeek-V3 swings hard technical punch at a lighter cost. ChatGPT still waves the flag for creative and explanatory slogging. Gemini 2.5 gulps down staggering amounts of stuff and chews on many data shapes at once in a way no other rig can touch. Qwen3 runs the table on translation and the crafting of what users see and poke. LLaMA stays the spine of open-source tinkering and private builds.

No single tool grabs the crown across every field. The right call swings entirely on the specific demands of the job at hand, including factors like the size of the input, the need for multimodal handling, the languages involved, and the available budget for inference costs. Two challenges stand out as particularly pressing. Current evaluation methods, especially widely used suites like MMLU and HumanEval, suffer from data contamination that undermines their reliability as true tests of model capability. At the same time, the gap between what benchmarks measure and what professional tasks actually require continues to complicate the selection process for practitioners. The real grind ahead is not about piling up ever-bigger models. It sits in nailing down how to test these things so the results actually tell something worth knowing. Until better yardsticks roll out, sticks that line up with how working professionals lean on these tools in their daily churn, scores and rankings will keep handing out a half-drawn map. What waits around the bend will most likely be rigs that can chew through longer, twistier strings of jobs on their own, making the hop from shooting back answers to actually moving the ball down the field.

References

[1] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. "LLaMA: Open and Efficient Foundation Language Models." 2023.

[2] Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Petrov, S., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., and Lazaridou, A. "Gemini: A Family of Highly Capable Multimodal Models." arXiv, 2312.11805, 2023.

[3] Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., and Viola, F. "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context." arXiv, 2403.05530, 2024.

[4] Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., Marris, L., Petulla, S., Gaffney, C., Aharoni, A., Lintz, N., Pais, T. C., Jacobsson, H., Szpektor, I., Jiang, N.-J., and Haridasan, K. "Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities." arXiv, 2507.06261, 2025.

DOI: 10.54254/2755-2721/2026.GL33655

[5] DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., and Luo, F. "DeepSeek-V3 Technical Report." arXiv, 2412.19437, 2024.

[6] Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., and Yang, J. "Qwen3 Technical Report." arXiv, 2505.09388, 2025.

[7] Awan, S. A., Khattak, M. A. K., Khan, A. A., Sathio, A. A., Alsayaydeh, J. A. J., Bacarra, R., Herawan, S. G., and Aziz, R. "Meta-Analysis of Large Language Models: Benchmarking DeepSeek-R1 Against ChatGPT, Gemini, Qwen, and LLaMA." Journal of Big Data, 2025.

[8] Jiang, Q., Gao, Z., and Karniadakis, G. E. "DeepSeek vs. ChatGPT vs. Claude: A Comparative Study for Scientific Computing and Scientific Machine Learning Tasks." Theoretical and Applied Mechanics Letters, 2025, p. 100583.

[9] Aydin, O., Karaarslan, E., Safa, E. F., and Bacanin, N. "Generative AI in Academic Writing: A Comparison of DeepSeek, Qwen, ChatGPT, Gemini, Llama, Mistral, and Gemma." arXiv, 2503.04765, 2025.

[10] Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., and Tang, Y. "A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development." IEEE/CAA Journal of Automatica Sinica, vol. 10, no. 5, 2023, pp. 1122-1136.