Human Interaction and Emerging Technologies (IHIET-AI), Vol. 201, 2026, 473–485
https://doi.org/10.54941/ahfe1007201

Hybrid Intelligence in the Innovation Process: Benchmark- and Utility-Based Selection of Proprietary Generative AI Models for Design Thinking in SMEs

Patrick Rupprecht and Isabel Rodenas

University of Applied Sciences for Management & Communication (FHWien der WKW), 1180 Wien, Währinger Gürtel 97, Austria

ABSTRACT

Small and medium-sized enterprises (SMEs) increasingly rely on generative AI to strengthen innovation processes while operating under tight resource and compliance constraints. Building on recent work on Hybrid Intelligence, this study presents a benchmark- and utility-based method to select enterprise-ready proprietary large language models (LLMs) for DesignThinking in EU-based SMEs. Using publicly available data from the Artificial Analysis Intelligence Index v3.0, the Hugging Face LMarena Leaderboard, and GDPR-aligned compliance criteria, we shortlist GPT-5.1, Gemini 3 Pro, Claude 4.5 Sonnet, and Magistral Medium 1.2. A two-stage utility analysis (unweighted and weighted) shows that Gemini 3 Pro consistently achieves the highest overall utility, particularly when reasoning quality, reliability, and speed are prioritised, followed by GPT-5.1. The analysis provides a transparent, replicable selection framework to support SMEs in adopting AI-assisted Design Thinking and outlines a practical foundation for orchestrating multiple models across innovation phases.

Keywords: Hybrid intelligence, Generative AI, Large language models, EU compliance, Benchmarking, Utility analysis, Design thinking, SMEs

INTRODUCTION

The accelerating development of generative artificial intelligence (AI) is reshaping innovation processes across industries and has become particularly relevant for small and medium-sized enterprises (SMEs). Operating under persistent constraints in financial resources, technical expertise, and innovation capacity (Ates & Bititci, 2011), SMEs increasingly rely on structured methodologies to guide decision-making and ensure responsible technology adoption. Design Thinking provides such a framework by combining user-centred inquiry, iterative experimentation, and interdisciplinary problem framing (Brown, 2008; Hasso Plattner Institute, 2023). At the same time, recent advances in generative AI have expanded the potential to augment analytical, creative, and interpretive tasks across all phases of Design Thinking. This development aligns with the emerging paradigm of Hybrid Intelligence, which emphasises the complementary strengths of human expertise and machine capabilities (Akata et al., 2020; Dellermann et al., 2019). Humans contribute

Received September 4, 2025; Revised November 28, 2025; Accepted December 10, 2025; Available online April 1, 2026

contextual judgment, ethical reasoning, and experience-based interpretation (Amershi et al., 2014; Jarrahi et al., 2022), while AI systems provide rapid content generation, large-scale knowledge access, and structured reasoning (Artificial Analysis, 2025; Stanford HAI, 2025). Empirical studies further indicate that generative AI can enhance productivity and improve output quality in knowledge-intensive work (Dell'Acqua et al., 2023; Dell'Acqua et al., 2025). Together, these factors position Hybrid Intelligence as a promising foundation for AI-assisted Design Thinking in SMEs.

However, despite growing availability of proprietary large language models (LLMs), SMEs face substantial uncertainty in selecting suitable systems. Model capabilities vary widely across reasoning, reliability, multimodal competence, latency, and operating costs. Additionally, European SMEs must comply with strict requirements regarding GDPR-aligned data processing, EU residency guarantees, and security certifications. Existing research provides limited guidance on how LLMs can be systematically evaluated and matched to the cognitive and procedural demands of Design Thinking.

To address this gap, this study develops a benchmark- and utility-based method for selecting proprietary, enterprise-ready LLMs for Design Thinking in EU-based SMEs. Drawing on publicly available performance data from the Artificial Analysis Intelligence Index v3.0 (Artificial Intelligence, 2025a) and the Hugging Face LMarena Leaderboard (Hugging Face, 2025), as well as GDPR- and security-related knockout criteria, the study identifies a shortlist of eligible models. A two-stage utility analysis (unweighted and weighted) evaluates their suitability for divergent and convergent Design Thinking tasks. The study aims to provide SMEs with a transparent, replicable selection framework that supports responsible and effective adoption of AI-assisted innovation workflows. Against this background, the research is guided by the following questions:

RQ1: Which proprietary LLMs meet EU compliance requirements and perform best on benchmark- and utility-based evaluation for Design Thinking in SMEs?

RQ2: How should these shortlisted models be evaluated - using criteria and utility analysis - to guide practical deployment in SME Design Thinking processes?

Hybrid Intelligence and Capabilities in the Design Thinking Innovation Process

Hybrid Intelligence describes cooperative systems in which humans and artificial intelligence (AI) complement each other's cognitive strengths to achieve superior outcomes compared with either acting alone (Akata et al., 2020; Dellermann et al., 2019; Rupprecht & Mayrhofer, 2024) (see figure 1). In innovation contexts, this paradigm provides a coherent lens through which to conceptualise how generative AI can augment but not replace human judgment, creativity, and interpretive competence. Humans contribute contextual awareness, tacit knowledge, ethical reasoning, and the ability to navigate ambiguity and user needs (Amershi et al., 2014; Jarrahi et al., 2022). AI systems, in contrast, provide large-scale pattern recognition, multimodal content generation, hypothesis exploration, and efficient processing of extensive or heterogeneous data (Artificial Analysis, 2025; Stanford HAI, 2025).

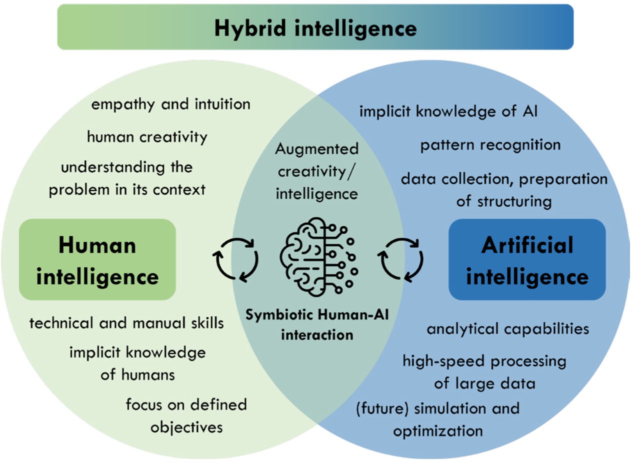

Hybrid intelligence
empathy and intuition
human creativity
understanding the problem in its context
Human intelligence
technical and manual skills
implicit knowledge of humans
focus on defined objectives
Augmented creativity/ intelligence
Symbiotic Human-AI interaction
implicit knowledge of AI
pattern recognition
data collection, preparation of structuring
Artificial intelligence
analytical capabilities
high-speed processing of large data
(future) simulation and optimization

Figure 1: Hybrid intelligence (Rupprecht & Mayrhofer, 2024, adapted from Dellermann et al., 2019).

These complementary strengths closely match the requirements of Design Thinking, which alternate between divergent and convergent modes of reasoning across phases such as Empathise, Define, Ideate, Prototype, and Test (Brown, 2008; Hasso Plattner Institute, 2023). Prior research indicates that effective Hybrid Intelligence additionally requires AI alignment and AI literacy, users must understand, supervise, and contextualise AI outputs to ensure that generated artefacts accurately reflect human intent and domain constraints (Dellermann et al., 2019). This is particularly relevant for SMEs, which often lack specialised AI departments and therefore benefit from systems that are reliable, transparent, and easy to supervise.

Building on these foundations, Rupprecht and Mayrhofer (2024) map human and AI capabilities to Design Thinking phases, showing how structured task allocation can increase both creative output and analytical depth. In early phases, humans lead user exploration and contextual interpretation, while AI supports data structuring and initial synthesis. During ideation and prototyping, AI enhances divergent creativity through rapid content generation and multimodal exploration. In later evaluative stages, AI contributes to structured comparison and predictive assessment, while humans retain responsibility for judgment, feasibility evaluation, and ethical considerations. This capability-based lens provides the conceptual rationale for the subsequent benchmarking and utility analysis. LLMs are therefore not evaluated in isolation but in relation to the cognitive demands and collaboration patterns of the Design Thinking process. The goal is to identify models that can be orchestrated effectively across phases, enabling SMEs to leverage Hybrid Intelligence in a targeted and outcome-driven manner.

SELECTION AND BENCHMARK-BASED EVALUATION OF PROPRIETARY LLMS FOR DESIGN THINKING IN SMES

The rapid expansion of proprietary large language models (LLMs) creates uncertainty for SMEs seeking enterprise-ready solutions for AI-assisted Design Thinking. To provide a transparent and replicable selection method, this study evaluates leading proprietary LLMs using benchmark performance and European compliance requirements based on publicly available data (Artificial Analysis, 2025; Hugging Face, 2025). All data reflect the status of 20 November 2025.

Benchmark Data Sources and Inclusion Criteria

Two independent benchmarking systems were used. First, the Artificial Analysis Intelligence (AAI) Index v3.0 provides a composite performance score across ten sub-benchmarks covering reasoning, knowledge, mathematical problem solving, programming, instruction following, long-context tasks, and agentic workflows (Artificial Analysis, 2025a). Second, the Hugging Face LMarena Leaderboard ranks models based on standardised evaluations (e.g., MMLU, GSM8K, ARC-Challenge, TruthfulQA) and community testing (Hugging Face, 2025).

Only models meeting the following inclusion criteria were considered:

(1) proprietary availability via commercial API;
(2) flagship status announced by the vendor;
(3) advanced reasoning features (e.g., chain-of-thought or tool use);
(4) readiness for enterprise deployment.

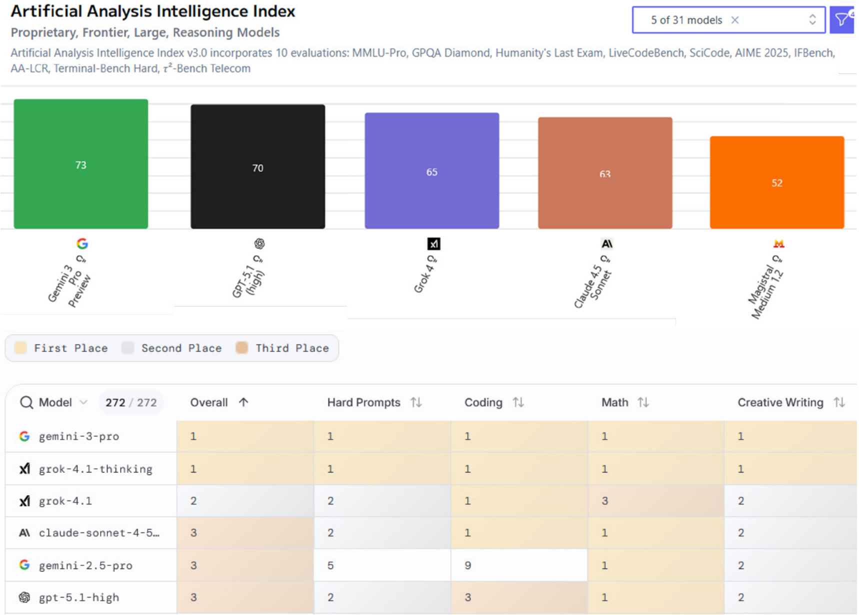
Open-source models (e.g., Llama, Qwen, Mistral open variants) were excluded due to missing enterprise guarantees regarding data protection, service-level agreements, and long-term security compliance.

Benchmark-Based Shortlisting

Models were shortlisted when they (a) ranked among the top proprietary systems in the AAI Index and (b) achieved Rank #1 on the LMarena Leaderboard. This produced the following candidates (see figure 2):

- **OpenAI GPT 5.1**, which represents the latest reasoning-oriented flagship from OpenAI and achieved the second highest overall score in the Artificial Analysis Index (OpenAI, 2025).
- **Anthropic Claude 4.5 Sonnet**, which achieved rank number 3 on the LMarena Leaderboard and rank number 4 on the AAI Leaderboard (Anthropic, 2025).
- **Google DeepMind Gemini 3 Pro**, the multimodal flagship and rank number 1 on LMarena and AAI Leaderboard (Google DeepMind, 2025).

- **Mistral AI Magistral Medium 1.2**, a proprietary EU hosted model with enterprise level API access, which is ranked on place 4 at AAI Index (Mistral AI, 2025).
- **xAI Grok 4**, which showed strong benchmark performance but was excluded because the vendor does not provide verifiable EU data residency options or security certifications (xAI, 2025).



Artificial Analysis Intelligence Index

Proprietary, Frontier, Large, Reasoning Models

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom

First Place Second Place Third Place

Model 272 / 272	Overall	Hard Prompts	Coding	Math	Creative Writing
gemini-3-pro	1	1	1	1	1
grok-4.1-thinking	1	1	1	1	1
grok-4.1	2	2	1	3	2
claude-sonnet-4-5…	3	2	1	1	2
gemini-2.5-pro	3	5	9	1	2
gpt-5.1-high	3	2	3	1	2

Figure 2: Benchmark results from *Artificial Analysis* (upper figure) and the *Hugging Face LMarena Leaderboard* (lower figure) (Artificial Analysis, 2025; Hugging Face, 2025). All data reflect the status of 20 November 2025.

Compliance and Data-Protection Assessment (KO Criteria)

For enterprise deployment in Austria and the broader European Union, compliance with data protection and information security regulations is a non-negotiable requirement. The assessment in this study therefore applies a set of knockout criteria derived from GDPR and enterprise security standards that determine whether a proprietary large language model can legally and operationally be used in SME environments:

(1) GDPR-compliant Data Processing Agreement (DPA) or Standard Contractual Clauses;
(2) EU data residency options;
(3) no training on customer data, or an enforceable opt-out;
(4) deletion and retention controls with audit logs;
(5) verified information-security certifications (ISO 27001 or SOC 2 Type II).

All compliance information is based on vendor documentation and validated public disclosures (OpenAI, 2025; Anthropic, 2025; Google DeepMind, 2025; Mistral AI, 2025; xAI, 2025).

Evaluation of vendor documentation confirms that GPT-5.1, Claude 4.5 Sonnet, Gemini 3 Pro, and Magistral Medium 1.2 satisfy all KO criteria, while Grok 4 does not. Four proprietary, enterprise-ready models proceed to the performance and utility evaluation:

- GPT-5.1: high reasoning capability; EU regional hosting; SOC 2 Type II compliant.
- Claude 4.5 Sonnet: strong reliability; ISO 27001 certified; EU processing.
- Gemini 3 Pro: multimodal flagship with leading benchmark results; EU hosting via GCP Belgium.
- Magistral Medium 1.2: EU-native model offering strong compliance and cost efficiency.

This shortlist forms the basis for the subsequent performance and economic evaluation through unweighted and weighted utility analyses.

Performance and Economic Evaluation (Utility Analysis)

Following the compliance assessment, the four eligible models were evaluated using a structured utility analysis that integrates benchmark performance and economic criteria. All performance data originate from the Artificial Analysis Leaderboard (Artificial Analysis, 2025), the Artificial Analysis Intelligence Index v3.0 (Artificial Analysis, 2025a), and the AA-Omniscience Index (Artificial Analysis, 2025b).

The utility analysis covers five evaluation criteria relevant for Design Thinking activities:

- General reasoning quality, based on the AAI composite score;
- Knowledge reliability and hallucination control, based on the AA-Omniscience Index;
- Latency, measured as time to first token;
- Output speed, measured in tokens per second;
- Price per token.

All benchmark values were normalised on a 1–5 scale; higher scores indicate stronger performance.

Unweighted Utility Analysis: The unweighted analysis (see table 1) assigns equal importance to all five criteria, representing balanced multi-phase Design Thinking usage.

Table 1: Unweighted utility analysis - results.

Criterion	GPT-5.1	Claude 4.5 Sonnet	Gemini 3 Pro	Magistral Medium 1.2
General quality and reasoning[1]	4	3	5	1
Knowledge reliability and hallucination tendency[2]	4	4	5	1
Latency (time to first token)[3]	1	4	4	5
Output speed (tokens per second)[3]	1	2	5	3
Price (normalised)[3]	3	1	2	5
Total points	13	14	21	15

Criteria weighting: all criteria equal (1 point each)

Rating scale: 1 = weakest performance, 5 = strongest performance

[1] Artificial Analysis (2025a). *Artificial Intelligence Index v3.0: Methodology and sub-benchmarks.*

[2] Artificial Analysis (2025b). *AA-Omniscience Index: Knowledge reliability and hallucination metric.*

[3] Artificial Analysis (2025). *Model performance metrics: Latency, output speed, pricing overview.*

Based on the recently updated benchmark data, results show that Gemini 3 Pro achieves the highest unweighted utility score (21), driven by outstanding reasoning, reliability and speed. Magistral Medium 1.2 follows due to excellent latency and cost performance (15). Claude 4.5 Sonnet and GPT-5.1 perform similarly (14 and 13), with GPT-5.1 penalised by latency and speed.

Weighted Utility Analysis: Because Design Thinking requires strong reasoning and factual reliability in convergent phases, a weighted analysis (see table 2) was applied:

- Quality & reasoning: 0.50,
- Knowledge reliability: 0.20
- Latency: 0.10, Output speed: 0.10, Price: 0.10

The weighting scheme reflects the relative impact of each criterion on Design Thinking performance. Quality & reasoning (0.50) receives the highest weight because convergent phases fundamentally rely on analytical coherence and structured problem solving. Knowledge reliability (0.20) is similarly critical, as low-hallucination outputs are essential for accurate framing and evaluation. Latency and output speed (each 0.10) affect user experience and iteration tempo but have limited influence on the substantive quality of results. Price (0.10) is weighted low because cost sensitivity varies strongly across SMEs, fluctuates frequently due to market dynamics, and typically does not determine the methodological quality of Design Thinking outcomes.

Table 2. Weighted utility analysis - results.

Criterion (Weight)	GPT-5.1	Claude 4.5 Sonnet	Gemini 3 Pro	Magistral Medium 1.2
General quality and logical reasoning (0.50)	4 × 0.50 = 2.00	3 × 0.50 = 1.50	5 × 0.50 = 2.50	1 × 0.50 = 0.50
Knowledge reliability and hallucination tendency (0.20)	4 × 0.20 = 0.80	4 × 0.20 = 0.80	5 × 0.20 = 1.00	1 × 0.20 = 0.20
Latency (0.10)	1 × 0.10 = 0.10	4 × 0.10 = 0.40	4 × 0.10 = 0.40	5 × 0.10 = 0.50
Output speed (0.10)	1 × 0.10 = 0.10	2 × 0.10 = 0.20	5 × 0.10 = 0.50	3 × 0.10 = 0.30
Price (0.10)	3 × 0.10 = 0.30	1 × 0.10 = 0.10	2 × 0.10 = 0.20	5 × 0.10 = 0.50
Total weighted score (0–5)	3.30	3.00	4.60	2.00

Criteria weighting: Quality and reasoning 0.50, knowledge reliability 0.20, latency 0.10, output speed 0.10, price 0.10.

Rating scale: 1 = weakest performance, 5 = strongest performance

Results indicate that Gemini 3 Pro achieves the highest weighted utility score (4.60), driven by superior quality & reasoning, strong knowledge reliability and low hallucination tendency, as well as favourable latency and throughput characteristics. GPT-5.1 (3.30) follows with strong quality & reasoning but lower operational efficiency due to weaker latency and output speed. Claude 4.5 Sonnet ranks third (3.00), offering stable quality & reasoning and solid knowledge reliability with a comparatively low hallucination tendency, but performing less favourably on speed and cost metrics. Magistral Medium 1.2 (2.00) provides competitive pricing and fast response times, yet is constrained by limited quality & reasoning and weaker knowledge reliability combined with a higher hallucination tendency.

Summary of utility findings, across both unweighted and weighted analyses:

- **Gemini 3 Pro** is the strongest all-round performer, particularly for iterative and time-critical Design Thinking phases, due to its superior quality & reasoning, strong knowledge reliability, and favourable latency and throughput.
- **GPT-5.1** excels in deep quality & reasoning and is well suited for analytical synthesis and problem framing, but is characterised by higher latency and lower throughput as well as comparatively higher operating costs.
- **Claude 4.5 Sonnet** provides high knowledge reliability and a low hallucination tendency, making it well suited for verification, risk-sensitive assessments, and regulatory or compliance-related tasks, despite weaker speed and cost efficiency.
- **Magistral Medium 1.2** delivers the strongest cost-performance ratio and fast response times, making it suitable for high-volume or routine tasks under EU-native hosting, although it is limited in quality & reasoning and knowledge reliability.

These findings provide a robust analytical foundation for phase-specific model allocation within Hybrid Intelligence-enabled Design Thinking workflows.

DISCUSSION

The results of the utility analysis demonstrate a differentiated performance landscape among the evaluated proprietary language models, which aligns closely with the requirements of Design Thinking. The process alternates between divergent phases, where fluency, speed and multimodal capability support broad exploration, and convergent phases that depend on accuracy, structured reasoning and factual reliability.

Across both unweighted and weighted analyses, Gemini 3 Pro emerges as the strongest general-purpose model for AI-supported Design Thinking. Its combination of high reasoning quality, exceptional speed, and competitive cost makes it particularly suitable for divergent ideation, rapid exploration, and iterative prototyping. These characteristics are consistent with findings that fast and fluent generative AI systems can increase creative output and accelerate early-stage innovation cycles (Dell'Acqua et al., 2025).

GPT-5.1 shows excellent deep reasoning performance and strong reliability, which makes it especially valuable for the convergent phases of Design Thinking. Tasks such as problem framing, thematic clustering, concept synthesis, and analytical evaluation benefit from the model's reasoning depth, even though its latency and throughput are lower than those of Gemini 3 Pro. As such, GPT-5.1 functions best as a precision model for high-stakes interpretive and evaluative scenarios.

Claude 4.5 Sonnet offers disciplined instruction following, which are advantageous in risk-sensitive settings, such as ethical assessment, regulatory evaluation or expert-level fact checking. Although its cost and speed performance are weaker, its controlled response behaviour makes it a dependable model for validation and high-trust tasks within later Design Thinking stages.

Magistral Medium 1.2 represents a cost-efficient EU-native option for SMEs operating under tight budget or strict data-governance constraints. While its reasoning performance falls short of the frontier models, it remains suitable for summarisation, documentation, basic ideation or other low-complexity tasks at scale.

Model Allocation Across Design Thinking Phases

The differentiated performance profiles may translate into a phase-sensitive allocation strategy for Design Thinking. While the frontier models - Gemini 3 Pro, GPT-5.1, and Claude 4.5 Sonnet - appear sufficiently capable across both divergent and convergent activities, their benchmarked characteristics suggest potential optimisation patterns rather than strict functional boundaries. Gemini 3 Pro may be particularly advantageous for divergent ideation and multimodal exploration due to its high quality & reasoning, strong knowledge reliability, and favourable latency and throughput. GPT-5.1, with its deep quality, reasoning and strong reliability, may be especially effective for convergent synthesis and analytical evaluation, while still being suitable for divergent work. Claude 4.5 Sonnet's high knowledge reliability and low hallucination tendency may make it a strong supporting tool for risk-sensitive assessments, provided it is used with robust human oversight. Magistral Medium 1.2, despite its limitations in quality & reasoning and knowledge reliability, may offer a cost-efficient EU-native option for routine or high-volume tasks with lower complexity requirements.

These observations indicate that SMEs may benefit from orchestrating a portfolio of complementary models instead of adopting a single-model strategy. Such an approach would be consistent with the Hybrid Intelligence paradigm, in which human expertise and differentiated AI capabilities interact dynamically to support varied cognitive demands across the innovation process (Dellermann et al., 2019).

Implications for Hybrid Intelligence in SMEs

The analysis reinforces three implications for Hybrid Intelligence-enabled innovation:

(1) Capability complementarity: Divergent and convergent phases profit from different model strengths, confirming that LLMs should be matched to the cognitive requirements of each design stage.
(2) Human-centred supervision: Even high-performing models require oversight to ensure contextual relevance, ethical alignment, and appropriate risk evaluation.
(3) Workflow integration: SMEs can improve innovation productivity by structuring workflows around explicit allocation rules that route tasks to the most suitable model. Overall, the findings provide a practical foundation for deploying multiple proprietary LLMs in Design Thinking and highlight how Hybrid Intelligence can strengthen creativity, analytical depth, and decision quality in SME innovation processes.

CONCLUSION AND OUTLOOK

This study examined how generative artificial intelligence may strengthen structured innovation methodologies, with a specific focus on Design Thinking in small and medium-sized enterprises. The research introduced an auditable and EU-compliant selection framework for proprietary large language models and illustrated how benchmark evidence and utility-based evaluation may support organizations in identifying suitable AI systems for different phases of the innovation process. The selection process integrated knockout criteria derived from GDPR and information-security requirements and combined these with performance benchmarks from the Artificial Analysis Intelligence Index v3.0 and the LMarena Leaderboard. This procedure yielded a shortlist comprising GPT-5.1, Gemini 3 Pro, Claude 4.5 Sonnet, and Magistral Medium 1.2, while Grok 4 was excluded due to missing EU residency guarantees and absent certifications.

The two-stage utility analysis suggests that Gemini 3 Pro may deliver the strongest overall performance in both unweighted and weighted evaluations, combining high quality & reasoning, strong knowledge reliability, favourable latency, and competitive operating costs. GPT-5.1 may remain particularly effective for analytical and synthesis-oriented tasks due to its deep quality & reasoning, while Claude 4.5 Sonnet may be advantageous in contexts requiring stable factual behaviour and controlled output characteristics. Magistral Medium 1.2 may offer a cost-efficient EU-hosted baseline for SMEs with constrained budgets or localised data requirements, despite its limitations in quality & reasoning and knowledge reliability.

Mapping these model profiles onto the Design Thinking framework indicates that the most effective approach for SMEs may not be to rely on a single frontier model. The three high-performing models - Gemini 3 Pro, GPT-5.1, and Claude 4.5 Sonnet - may all be suitable for both divergent and convergent phases due to their strong results in quality & reasoning and knowledge reliability. SMEs may still optimise performance by leveraging their differing operational characteristics: for example, speed and throughput in the case of Gemini 3 Pro, deep analytical reasoning with GPT-5.1, or stable and controlled response behaviour with Claude 4.5 Sonnet. Magistral Medium 1.2 may remain a viable option for documentation-heavy, routine, or lower-complexity tasks where cost efficiency and latency are more critical than advanced reasoning capability.

Several limitations may affect the generalizability of the findings. The analysis draws on publicly available benchmark data and vendor documentation, which may not fully reflect model performance in specific organizational environments. The utility weighting scheme reflects the priorities of Design Thinking processes and may differ across industries or under alternative constraints such as cost sensitivity in SMEs. Furthermore, the focus on proprietary frontier models excludes open-source alternatives that may be viable for organizations with strong internal governance and technical capabilities.

Future work includes the development of an AI-assisted Design Thinking chatbot that can orchestrate multiple proprietary models across different task types. In a subsequent evaluation study with innovation experts from SMEs, a complete Design Thinking process will be executed using this system, enabling assessment of quality & reasoning, human–AI interaction, and the usability of the generated outputs. To ensure an unbiased comparison, all four shortlisted models will be evaluated in a blind-test setting. Such a study may provide deeper insights into how SMEs may practically adopt and supervise orchestrated AI support within collaborative innovation workflows.

By combining benchmark evidence, compliance validation, and utility-oriented model selection, this study provides an analytical foundation for SMEs intending to implement AI-assisted innovation processes. The findings suggest that Hybrid Intelligence may offer a feasible and ethically aligned path for strengthening creativity, analytical depth, and decision quality in SME-driven innovation.

Ethics Statement

This study did not involve human or animal subjects. All analyses were conducted using publicly available benchmark data and vendor documentation. No personal or sensitive data were collected or processed. Ethical considerations related to transparency and data protection were addressed through GDPR-aligned screening and the verification of ISO 27001 and SOC 2 compliance claims where applicable.

REFERENCES

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F. and Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8), 18–28. https://doi.org/10.1109/MC.2020.2996587

Amershi, S., Cakmak, M., Knox, W. B. and Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

Anthropic (2025). Claude 4.5 Sonnet. https://www.anthropic.com/news/claude-sonnet-4-5

Artificial Analysis (2025). Comparison of models: Intelligence, performance & price analysis. https://artificialanalysis.ai/models

Artificial Analysis (2025a). Intelligence (AAI) Index v3.0. https://artificialanalysis.ai/methodology/intelligence-benchmarking#artificial-analysis-intelligence-index

Artificial Analysis (2025b). Knowledge and hallucination – AA-Omniscience. https://artificialanalysis.ai/methodology/intelligence-benchmarking#aa-omniscience

Ates, A. and Bititci, U. (2011). Change process: A key enabler for building resilient SMEs. *International Journal of Production Research*, 49(18), 5601–5618. https://doi.org/10.1080/00207543.2011.563825

Brown, T. (2008). Design thinking. *Harvard Business Review*. https://hbr.org/2008/06/design-thinking

Dell'Acqua, F., Ayoubi, C., Lifshitz-Assaf, H., Sadun, R., Mollick, E. R., Mollick, L., Han, Y., Goldman, J., Nair, H., Taub, S. and Lakhani, K. R. (2025). The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise. *Harvard Business School Working Paper* 25-043. https://ssrn.com/abstract=5188231

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S. and Ebel, P. (2019). The future of human–AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems. In T. X. Bui (ed.), *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences*. University of Hawai'i at Mānoa.

Dellermann, D., Ebel, P., Söllner, M. and Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 61(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

Dell'Acqua, F., McFowland, E. III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School*

Technology & Operations Management Unit Working Paper 24-013. https://doi.org/10.2139/ssrn.4573321

Google DeepMind (2025). Gemini 3 Pro. https://deepmind.google/models/gemini/pro/

Hasso Plattner Institute of Design at Stanford University (2023). Design Thinking process. https://dschool.stanford.edu

Hugging Face (2025). LMarena Leaderboard. https://huggingface.co/spaces/lmarena-ai/lmarena-leaderboard

Jarrahi, M. H., Lutz, C. and Newlands, G. (2022). Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. *Big Data & Society*, 9(2). https://doi.org/10.1177/20539517221142824

Mistral AI (2025). Magistral Medium 1.2 Model. https://docs.mistral.ai/models/magistral-medium-1-2-25-09

OpenAI (2025). GPT-5.1. https://openai.com/de-DE/index/gpt-5-1

Rupprecht, P. and Mayrhofer, W. (2024). Hybrid intelligence – An approach towards the symbiosis of artificial and human creativity and interaction in the design and innovation process in SMEs. In *Creativity, Innovation and Entrepreneurship*. AHFE International. https://doi.org/10.54941/ahfe1004718

Stanford HAI (2025). *2025 AI Index Report*. https://hai.stanford.edu/ai-index/2025-ai-index-report

xAI (2025). Grok 4. https://docs.x.ai/docs