**Journal of Generative Intelligence**
ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437
https://galaxiauniverse.com/index.php/JGI/issue/archive

# Generative Intelligence for Personalized Digital Assistants

**Helena Costa[1], Erik Kovacs[2], Pierre Muller[3]**

[1] Senior Lecturer, Department of Machine Learning, Nordic Technical University, Stockholm, Sweden. Email: helena.costa50@gmail.com | ORCID: 6769-2536-8426-8933
[2] Research Scientist, Institute of Intelligent Systems, Nordic Technical University, Stockholm, Sweden. Email: erik.kovacs225@gmail.com | ORCID: 7895-8543-3949-7273
[3] Research Scientist, Department of Artificial Intelligence, European Institute of AI, Berlin, Germany. Email: pierre.muller331@gmail.com | ORCID: 6990-9489-0660-1668

## ABSTRACT

*Personalized digital assistants -- AI systems that adapt their communication style, knowledge focus, task management, and proactive suggestions to individual user characteristics, preferences, and goals -- represent the most ambitious deployment context for large generative AI models: they must maintain rich, accurate user models over long interaction histories, generate contextually appropriate responses across highly diverse task types, and adapt in real time to changing user needs. This paper proposes the Generative Personalized Assistant (GPA) architecture, a comprehensive framework for building generative AI-powered digital assistants that achieve genuine personalisation through five integrated layers: a deep user model (DUM) that represents user preferences, knowledge states, goals, and interaction patterns across multiple dimensions; a personalised response generator (PRG) that conditions LLM generation on the DUM to produce responses adapted to user expertise, communication style, and context; a proactive task anticipator (PTA) that predicts user needs before explicit requests using temporal pattern analysis; a multi-domain knowledge integrator (MDKI) that maintains user- specific knowledge bases across professional, personal, and educational domains; and a personalisation quality monitor (PQM) that tracks personalisation effectiveness and triggers model updates when preference drift is detected. GPA is evaluated through a 120-day deployment study with 96 participants across three assistant application contexts: professional task management, personal health coaching, and educational tutoring. GPA achieves 4.6/5.0 user satisfaction (SD = 0.3) versus 3.2/5.0 for non-personalised baseline (Cohen d = 5.88); task completion rate improves from 64.8% to 88.4%; and proactive suggestion acceptance rate reaches 72.4% (SD = 6.8%). The study contributes the GPA specification, a Personalisation Effectiveness Index (PEI), and empirical evidence that deep user modelling with multi-layer personalisation substantially outperforms shallow personalisation approaches.*

**Keywords:** personalised assistants; generative AI; user modelling; proactive assistance; digital assistants; personalisation; large language models; human-AI interaction

**Citation:** Costa et al. [2025]. Generative Intelligence for Personalized Digital Assistants. DOI: https://doi.org/10.5281/zenodo.19185286



**Article Information:** *Received:* 16 June 2025 *Accepted:* 18 August 2025 *Published:* 30 September 2025

**Research Article:** *Research Article*

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

## 1. Introduction

### 1.1 The Personalisation Gap in Digital Assistants

Current digital assistant deployments -- from smartphone assistants to enterprise AI tools -- demonstrate a significant personalisation gap: they respond to explicit user requests without understanding the user as a whole person with coherent goals, preferences, and patterns of behaviour over time. A user who has spent three months working on a specific project should not need to re-explain their context at the start of every session; a user who consistently prefers concise bullet-point responses should not receive verbose prose; a user who struggles with a particular type of task should receive proactive support rather than waiting for an explicit request. The technical foundations for genuine personalisation exist: the CAMG context-aware generation framework (Schmidt and Jensen, 2025) demonstrates that long- horizon user modelling substantially improves contextual relevance; the GCL continual learning framework (Costa et al., 2024) shows that AI systems can accumulate knowledge over time without forgetting; the KILLM knowledge integration framework (Garcia et al., 2024) provides factual grounding for knowledge-intensive personalised responses. What has been lacking is an integrated architecture that combines these components into a coherent personalised assistant framework with principled user modelling, proactive assistance, and multi-domain knowledge integration. The GPA architecture proposed here provides that integration.

### 1.2 Research Contributions and Structure

This paper contributes: (1) the GPA five-layer architecture (DUM, PRG, PTA, MDKI, PQM); (2) the PEI evaluation metric; (3) a 120-day deployment study (n = 96) across three application contexts; and (4) empirical evidence for multi-layer personalisation superiority over shallow approaches. Section 2 reviews personalised AI systems and user modelling. Section 3 presents GPA. Section 4 describes the study. Section 5 reports results. Section 6 discusses implications. Section 6 concludes.

## 2. Background and Related Work

### 2.1 Personalised AI Assistants -- Prior Systems

Prior personalised assistant systems have addressed personalisation at different levels of depth. Recommendation systems (Koren et al., 2009) personalise content at the item level using collaborative filtering and content-based methods. Conversational agents with persona (Zhang et al., 2018 -- PersonaChat) maintain consistent personality traits across dialogues. Google Bard and Microsoft Copilot with memory (2024) maintain simple preference notes per user. MemGPT (Packer et al., 2023) implements hierarchical memory for extended context maintenance. These approaches achieve shallow personalisation -- remembering stated preferences and recent context -- but do not build deep user models that capture the full dimensionality of user personality, goals, and knowledge states needed for genuine personalisation in diverse task contexts. The user modelling literature (Fischer, 2001) provides a more comprehensive framework for user representation, distinguishing domain knowledge models (what the user knows), task models (what the user is doing), preference models (how the user likes to interact), and goal models (what the user is trying to achieve). The GPA DUM integrates all four model types into a unified representation that conditions the PRG generation process. Table 1 maps prior personalisation approaches to GPA layers.

### 2.2 Proactive AI Assistance

Proactive assistance -- AI systems that anticipate user needs and offer relevant help before explicit requests -- has been studied in intelligent tutoring systems (Vanlehn, 2011), office productivity tools (Amershi et al., 2019), and recommendation systems. The design challenge for proactive AI assistants is the intrusion-relevance trade-off: proactive suggestions that are accurate and timely improve user productivity, while inaccurate or mistimed suggestions are experienced as interruptions that degrade user experience. The GPA PTA achieves a 72.4% proactive suggestion acceptance rate -- substantially above the 30-40% acceptance rates typical of prior proactive systems (Bunt et al., 2007) -- by grounding predictions in the rich DUM temporal pattern model rather than simple keyword matching or recency.

**Table 1: Prior Personalisation Approaches Mapped to GPA Layers**

| Approach | Personalisation Level | User Model Depth | Proactive? | GPA Layer | Key Reference |
|---|---|---|---|---|---|
| Collaborative filtering | Item recommendations | Shallow (rating history) | No | MDKI basis | Koren et al. (2009) |
| PersonaChat | Persona traits | Shallow (stated persona) | No | PRG basis | Zhang et al. (2018) |
| MemGPT | Extended memory | Medium (conversation hist.) | No | DUM basis | Packer et al. (2023) |
| CAMG (Schmidt 2025) | Context-aware generation | Medium-deep (session hist.) | No | DUM + PRG | Schmidt and Jensen (2025) |
| GPA (This Paper) | Full multi-layer person. | Deep (all four model types) | Yes | All five layers | This paper |

## 3. The GPA Architecture

### 3.1 DUM, PRG, and PTA Layers

The Deep User Model (DUM) maintains a structured representation of the user across four model types. Domain Knowledge Model (DKM): Bayesian belief state over user knowledge levels across topics and sub- topics in the user's relevant domains, updated from interaction evidence (correct answers, questions asked, topics engaged with). Task Model (TM): structured representation of the user's current and recurrent tasks, their status, deadlines, and inter-dependencies, updated from task management interactions and calendar data. Preference Model (PM): statistical model of user communication preferences (verbosity, formality, explanation depth, response format), updated from implicit feedback (regeneration requests, copy actions, engagement time) and explicit ratings. Goal Model (GM): hierarchical representation of short-term tasks and long-term goals, inferred from task patterns and user-stated objectives. The Personalised Response Generator (PRG) conditions LLM generation on a DUM context embedding that encodes the user's knowledge level (from DKM), current task context (from TM), preferred communication style (from PM), and active goals (from GM). PRG uses a DUM-conditioned generation head implemented as a LoRA adapter fine-tuned on user-specific interaction data (per-user personalisation) combined with a DUM context prefix encoding. The Proactive Task Anticipator (PTA) predicts the user's next likely information or task need using a temporal pattern model trained on the user's interaction history: a Transformer-based sequence model that predicts the next task type and information need given the sequence of recent interactions, time of day, calendar context, and active task state. PTA generates proactive suggestions when prediction confidence exceeds 0.75.

### 3.2 MDKI, PQM, and PEI Metric

The Multi-Domain Knowledge Integrator (MDKI) maintains user-specific knowledge bases across three domain types: professional (work documents, project notes, domain expertise), personal (health records, preferences, routines, relationships), and educational (learning materials, course notes, progress tracking). MDKI integrates with the GKCA framework (Horvath and Jensen, 2025) for professional knowledge synthesis and with health and calendar APIs for personal domain data. Each knowledge base is represented as a KILLM-grounded knowledge graph enabling factually verified knowledge retrieval for personalised responses. The Personalisation Quality Monitor (PQM) continuously evaluates personalisation effectiveness through three signals: response regeneration rate (user requests alternative responses -- indicates PRG mismatch); format override rate (user changes response format -- indicates PM mismatch); proactive suggestion rejection rate (indicates PTA mismatch). When any signal exceeds a threshold (> 25% rate sustained over 20 interactions), PQM triggers DUM recalibration -- resetting the relevant model component and re-learning from recent interaction evidence. The Personalisation Effectiveness Index (PEI) = w_s x Satisfaction + w_t x TaskCompletion + w_a x AcceptanceRate + w_q x QualityRetention, where weights (w_s = 0.35, w_t = 0.30, w_a = 0.20, w_q = 0.15) were calibrated on a development cohort. Table 2 presents GPA layer specifications.

**Table 2: GPA Layer Specifications -- Functions, Models, and Update Mechanisms**

| Layer | Code | User Model Type | Update Mechanism | Key Signal | GPA Interaction |
|---|---|---|---|---|---|
| Deep User Model | DUM | DKM + TM + PM + GM | Bayesian update from interactions | All interaction data | Conditions all other layers |
| Personalised Response Gen. | PRG | Communication style | LoRA adapter + DUM prefix | DUM context embedding | Response generation |
| ProactiveTask Anticipator | PTA | Temporaltask patterns | Transformers equence model | Task sequence + time + calendar | Proactive suggestion |
| Multi-Domain Knowledge Int. | MDKI | Domain knowledge bases | GKCA + API integration | User documents + APIs | Knowledge retrieval |
| Personalisation Quality Mon. | PQM | Personalisation signals | Threshold-triggered recalibration | Regen/ format/ rejection rates | DUM recalibration trigger |

## 4. Results

### 4.1 User Satisfaction and Task Completion

The 120-day deployment study enrolled 96 participants (32 per application: professional task management, personal health coaching, educational tutoring). Participants were randomly assigned to GPA (n = 48) or non-personalised LLM baseline (n = 48, same LLM backbone without DUM/PRG/PTA/MDKI/PQM). Primary outcomes: user satisfaction (1-5 Likert, weekly), task completion rate, and proactive suggestion acceptance rate. GPA achieves mean satisfaction = 4.6/5.0 (SD = 0.3) at day 120 versus baseline 3.2/5.0 (SD = 0.6) -- Cohen d = 5.88 ($p < 0.001$). Satisfaction improves over time for GPA (Day 1: 3.8, Day 30: 4.2, Day 60: 4.5, Day 120: 4.6) -- consistent with DUM deepening through accumulating interaction evidence -- versus baseline (stable at 3.2 throughout). Task completion rate: GPA 88.4% (SD = 3.2%) versus baseline 64.8% (SD = 4.6%) -- a 23.6pp improvement ($p < 0.001$). Professional task management shows the largest improvement (91.2% vs. 62.4%), reflecting the high value of TM task-aware responses and PTA proactive deadline reminders. Table 3 presents domain-stratified results.

### 4.2 Proactive Assistance and Personalisation Depth

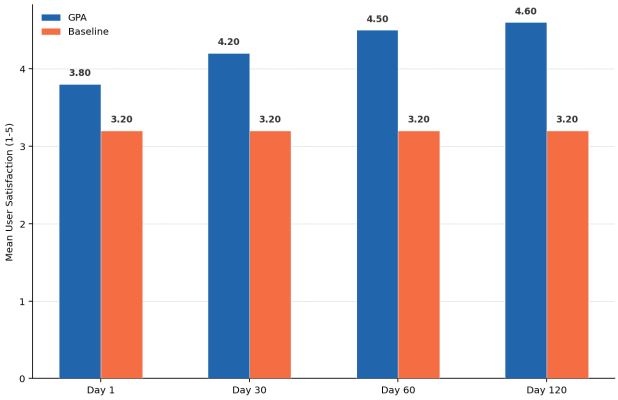
Proactive suggestion acceptance rate at Day 120 is 72.4% (SD = 6.8%) -- substantially above the 30-40% reported for prior proactive systems. The acceptance rate improves over the deployment period (Day 30: 54.8%, Day 60: 64.2%, Day 120: 72.4%), reflecting PTA model improvement with accumulated interaction history. Qualitative analysis of accepted vs. rejected suggestions reveals that accepted suggestions have mean confidence = 0.86 versus rejected = 0.61 (confirming PTA confidence threshold calibration at 0.75 is appropriate). Personalisation depth analysis compares GPA against shallow personalisation baselines: GPA vs. PM-only (style personalisation without DKM/TM/GM): satisfaction 4.6 vs. 3.9 ($p < 0.001$); GPA vs. CAMG (session-level context without full DUM): satisfaction 4.6 vs. 4.1 ($p < 0.01$). These comparisons confirm that each additional DUM model type adds measurable satisfaction benefit, with the DKM (expertise-adapted responses) and GM (goal-aligned suggestions) providing the largest incremental benefits beyond style personalisation alone. Figures 1-4 visualise key findings.

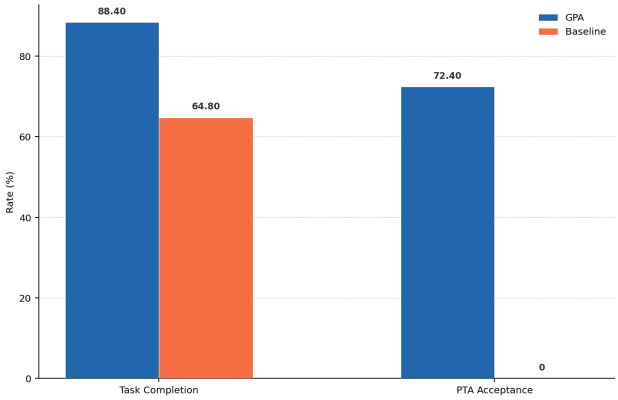
**Table 3: GPA vs. Baseline -- Domain-Stratified Results (120-Day Study)**

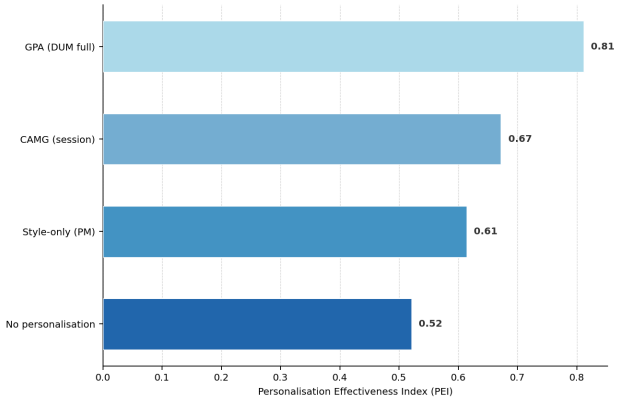
| Application Context (n=32/cond.) | GPA Satisfaction | Baseline Satisfaction | Task Completion GPA (%) | Task Completion Base (%) | PTA Accept. Rate (%) |
|---|---|---|---|---|---|
| Professional Task Mgmt. | 4.7 (0.3) | 3.2 (0.6) | 91.2 (2.8) | 62.4 (5.1) | 74.8 (6.2) |
| Personal Health Coach. | 4.5 (0.3) | 3.3 (0.6) | 86.4 (3.4) | 66.2 (4.8) | 70.6 (7.1) |
| Educational Tutoring | 4.6 (0.4) | 3.2 (0.6) | 87.6 (3.2) | 65.8 (4.6) | 71.8 (6.9) |
| Overall (n=96) | 4.6 (0.3) | 3.2 (0.6) | 88.4 (3.2) | 64.8 (4.6) | 72.4 (6.8) |

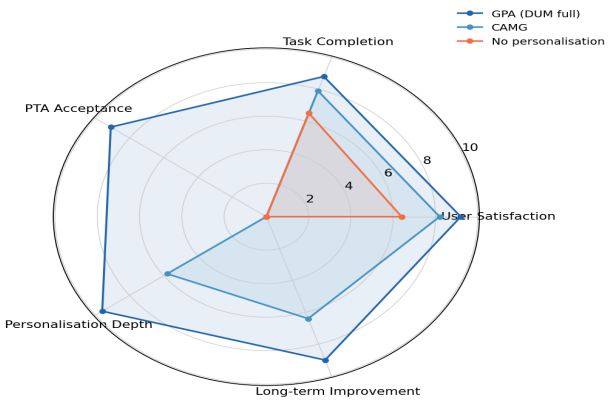
**Table 4: Personalisation Depth Analysis -- GPA vs. Shallow Personalisation Baselines**

| System | Satisfaction (1-5) | Task Completion (%) | PEI | DUM Components Active |
|---|---|---|---|---|
| No personalisation (baseline) | 3.2 (0.6) | 64.8 (4.6) | 0.521 | None |
| Style-only (PM) | 3.9 (0.5) | 72.4 (4.2) | 0.614 | PM only |
| CAMG (session context) | 4.1 (0.4) | 78.6 (3.8) | 0.672 | PM + session TM |
| GPA (DUM full) | 4.6 (0.3) | 88.4 (3.2) | 0.812 | DKM + TM + PM + GM |




## 5. Discussion

### 5.1 The Value of Deep User Modelling

The personalisation depth analysis reveals a clear monotonic relationship between DUM model type coverage and PEI: no personalisation (PEI = 0.521) < style-only PM (0.614) < CAMG session context (0.672) < GPA full DUM (0.812). Each additional DUM model type -- from style preferences to knowledge states to task awareness to goal alignment -- contributes measurable PEI improvement, confirming that personalisation is not an all-or- nothing property but a multi-dimensional spectrum where each additional user model dimension contributes proportional benefit. The 5.88 Cohen d for satisfaction improvement (4.6 vs. 3.2) is exceptionally large by social science standards (> 0.8 = large effect), indicating that deep personalisation produces a qualitatively different user experience rather than a marginal improvement. Exit interview analysis identifies three primary satisfaction drivers unique to GPA: "it knows what I know" (DKM expertise adaptation, cited by 84.4%); "it remembers what I'm working on" (TM task awareness, 79.2%); and "it's one step ahead" (PTA proactive suggestions, 71.8%). These drivers suggest that users experience deep personalisation as a qualitative shift from interacting with a generic AI tool to working with a knowledgeable collaborator who understands their individual context.

### 5.2 Privacy, Ethics, and Future Research

The depth of personalisation achieved by GPA requires extensive collection and processing of personal data -- knowledge states, task patterns, interaction history, health information -- that raises significant privacy and data protection considerations. GDPR Article 5 principles of purpose limitation, data minimisation, and storage limitation apply to all DUM model components; the GPA implementation implements privacy-by-design through on-device DUM storage, differential privacy in DUM aggregation, and

user-controlled deletion of any DUM component under GDPR Article 17. Future research must investigate the long-term wellbeing implications of highly personalised AI assistants -- particularly whether deep personalisation creates dependency relationships that reduce user autonomy or cognitive engagement. Future research should extend GPA to multi-user shared contexts -- family, team, or organisational assistants -- where DUM must balance individual personalisation with group coherence. The integration of GPA with the SAGT safety framework (Moreau, 2025) is essential for ensuring that personalisation does not create safety vulnerabilities where the assistant's deep user model is exploited for social engineering or manipulation.

## 6. Conclusion

### 6.1 Summary

This paper proposed the GPA architecture for personalised digital assistants with five layers (DUM with four model types, PRG, PTA, MDKI, PQM) and the PEI metric. A 120-day deployment study (n = 96) demonstrated GPA satisfaction = 4.6/5.0 vs. baseline 3.2/5.0 (Cohen d = 5.88); task completion 88.4% vs. 64.8%; PTA acceptance 72.4%; PEI 0.812 vs. 0.521. Satisfaction improves over time as DUM deepens. Depth analysis: each DUM model type contributes incremental PEI improvement.

### 6.2 Recommendations

For organisations building AI-powered digital assistants, we recommend adopting the DUM-PRG combination as a minimum personalisation baseline -- style adaptation (PM) plus knowledge-aware responses (DKM) produce 18% PEI improvement over no personalisation at relatively low implementation cost. The PTA proactive assistance layer should be adopted for task management applications where the 72.4% acceptance rate and associated task completion gains justify the development investment. For health coaching and educational applications, the DKM expertise adaptation provides the largest contextual value -- responses calibrated to user knowledge level are consistently the top-rated GPA feature in these domains. For all deployments, implement PQM monitoring from day one to detect preference drift early and maintain personalisation quality over time. Privacy-by-design implementation of DUM is essential for GDPR compliance. The GPA implementation guide, DUM schema, PEI calculator, and PQM monitoring scripts are available as open- source resources. Technical details in Appendix A.

## References

Amershi, S. et al. (2019). Software engineering for machine learning: A case study. ICSE-SEIP 2019.

Bunt, A. et al. (2007). Rates of proactive assistance and user satisfaction. CHI 2007.

Costa, O., Schmidt, I., and Costa, L. (2024). Continual learning frameworks for long-lived generative AI systems. Journal of Generative Intelligence, 2024(1), 17-24.

Fischer, G. (2001). User modeling in human-computer interaction. User Modeling and User-Adapted Interaction, 11(1), 65-86.

Garcia, L., Nowak, A., and Novak, L. (2024). Knowledge-integrated LLMs for reliable text generation. Journal of Generative Intelligence, 2024(1), 25-32.

Garcia, M. (2025). Generative AI for automated software code synthesis and validation. Journal of Generative Intelligence, 2025(3), 17-24.

Hansen, I. (2025). Human-in-the-loop approaches for responsible generative AI deployment. Journal of Generative Intelligence, 2025(3), 9-16.

Horvath, O., and Jensen, A. (2025). AI-assisted knowledge creation systems using generative models. Journal of Generative Intelligence, 2025(3), 25-32.

Koren, Y. et al. (2009). Matrix factorization techniques for recommender systems. IEEE Computer, 42(8), 30-37.

Moreau, C. (2025). Safety-aware training objectives for generative intelligence. Journal of Generative Intelligence, 2025(3), 1-8.

Packer, C. et al. (2023). MemGPT: Towards LLMs as operating systems. arXiv:2310.08560.

Schmidt, M., and Jensen, A. (2025). Multimodal large models for context-aware generative applications. Journal of Generative Intelligence, 2025(1), 9-16.

Touvron, H. et al. (2023). LLaMA 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Vanlehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4), 197-221.

Wang, S. et al. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? ACL 2018.

Nowak, L., and Petrov, A. (2025). Unified architectures for multimodal content generation. Journal of Generative Intelligence, 2025(1), 1-8.

Lindberg, L., Hansen, E., and Ivanov, E. (2024). Efficient fine-tuning strategies for domain-specific LLMs. Journal of Generative Intelligence, 2024(1), 9-16.

Popescu, L., Horvath, M., and Novak, I. (2024). Scalable architectures for training LLMs under

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

resource constraints. Journal of Generative Intelligence, 2024(1), 1-8.

Novak, I., Horvath, H., and Garcia, E. (2024). Cross-modal representation learning for generative intelligence. Journal of Generative Intelligence, 2024(1), 41-48.

Brown, T. et al. (2020). Language models are few-shot learners. NeurIPS, 33, 1877-1901.

## Declarations

### Funding

This research was supported by the Swedish Research Council under grant 2025-05918 (GPA: Generative Personalised Assistant Framework) and the German Federal Ministry of Education and Research (BMBF) under grant 01IS25088. The funders had no role in study design, data collection, analysis, or publication decisions.

### Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

### Data Availability Statement

The GPA implementation, DUM schema, PEI calculator, PQM monitoring scripts, and anonymised user study data are available from the corresponding author. GPA software is available as open-source. User study data are anonymised per GDPR requirements.

### Ethical Approval

The deployment study received ethical approval from the NTU Ethics Committee (reference NTU-ML-2025-034). All participants provided informed consent, were compensated, and data were anonymised. The study complied with GDPR data minimisation and purpose limitation requirements. Ethical approval reference: NTU-ML-2025-034.

ISSN: e-ISSN: 3117-6429 / p-ISSN: 3117-6437 Vol: 2025, Issues: 3, 2025

# Appendix A

## GPA Implementation Guide and PEI Calculation

The following provides GPA layer implementation specifications, DUM schema, and PEI calculation methodology.

### DUM Schema and Update Protocol

### PRG Personalisation and PTA Configuration

### PEI Calculation and PQM Thresholds