Decoupling Intelligence from Identity: The CPAP Protocol, the AI Ledger, and a Zero-Trust Middleware for Sovereign Enterprise LLMs

Nitin Lodha
Chitrangana
nitin@chitrangana.com

5 August 2026 (revised)

Revision Notice. This paper significantly revises and expands upon the preliminary version published at SSRN: Lodha, Nitin, *A Zero-Trust Hybrid Architecture for Enterprise LLM Deployment* (August 5, 2026). Available at SSRN: https://ssrn.com/abstract=7236658 or http://dx.doi.org/10.2139/ssrn.7236658. The present version includes formal design rationale, corrected algorithms, expanded related work, and a reproducible benchmark suite.

Abstract

Enterprise adoption of large language model (LLM) APIs creates two acute operational challenges: (1) data sovereignty—how to use cloud AI without exposing confidential client data; and (2) API drift—how to maintain stable AI behaviour when cloud providers update models without notice. This technical report describes a zero-trust hybrid middleware architecture deployed across multiple enterprise clients under active client confidentiality agreements. The architecture comprises four composable components: the Capability-Preserving Anonymization Protocol (CPAP), a reversible semantic masking layer; a Dynamic Router implementing the Routing Score Matrix (RSM) with a circuit-breaker penalty box; a Local Verification Council of deterministic guardrails; and an AI Ledger that accumulates operational intelligence to counteract model drift. We present the system design, complete architectural specifications, and reproducible benchmark evaluations on a synthetic enterprise dataset ($n = 100$ prompts, seed = 42). CPAP achieves 100% masking accuracy and perfect reversibility at 10,897 prompts/second. The Dynamic Router delivers 67.1% cost reduction against an always-GPT-4o baseline while retaining 90.4% of peak quality. The AI Ledger reduces drift-induced error rates by 70.0% through accumulated corrections. All components are released as open-source reference implementations under the MIT License.

Keywords: enterprise LLM, data sovereignty, API drift, reversible semantic masking, LLM routing, zero-trust AI, prompt governance.

NDA Disclosure. Empirical observations are drawn from active enterprise deployments between October 2025 and July 2026. Due to binding confidentiality agreements, specific organizational identities, financial figures, and proprietary data schemas are not disclosed. All quantitative benchmark results in this report are reproducible using the open-source synthetic dataset included in the repository.

1 Introduction

Enterprise AI deployments face a fundamental paradox. Cloud LLM APIs offer state-of-the-art language capabilities, but using them requires transmitting enterprise data—often regulated, confidential, or proprietary—to external infrastructure operated by third parties. Simultaneously, cloud providers silently update their models, causing enterprise workflows calibrated to one model's behaviour to degrade without warning, a phenomenon we term API drift.

This technical report describes a middleware architecture developed through active enterprise deployments. We present:

(1) the complete architectural specification of four interoperating components;

(2) a reproducible synthetic benchmark evaluating each component independently; and

(3) a reference open-source implementation sufficient for independent replication.

The architecture is built on a single design principle: decouple intelligence from identity. Cloud models provide reasoning capability; local infrastructure retains data sovereignty.

1.1 Problem Statement

Definition 1 (Data Sovereignty Problem)**.** *Let D be an enterprise document containing sensitive entities $E = \{e_1, \ldots, e_n\}$. An enterprise must obtain LLM inference on D without exposing any e_i to the cloud provider.*

Definition 2 (API Drift Problem)**.** *Let f_t denote the cloud LLM function at time t. An enterprise's workflow W is calibrated to f_t. At time $t' > t$, the provider silently deploys $f_{t'} \neq f_t$. The enterprise must maintain workflow behaviour without re-engineering W for each model change.*

1.2 Contributions

This paper makes the following contributions:

(1) **CPAP:** a provably reversible semantic masking protocol with HMAC-SHA256 token derivation and information-theoretic leakage bounds;

(2) **Dynamic Router:** RSM-based routing with circuit-breaker quarantine and cost-accuracy optimisation matrix;

(3) **AI Ledger:** instruction accumulation framework with formal update rule and weekly consolidation;

(4) **Local Verification Council:** deterministic Boolean guardrails for pre/post LLM validation;

(5) a synthetic benchmark dataset (100 prompts, seed = 42, publicly available); and

(6) an open-source reference implementation under MIT License at `https://github.com/chitrasource/chitrangana-ai-middleware`.

2 Related Work

2.1 Privacy-Preserving LLM Inference

Protecting data during cloud LLM inference is an active research area. Homomorphic encryption (HE) approaches [1] offer cryptographic security guarantees but impose 10^2–$10^4\times$ computational overhead, making real-time inference infeasible. Differential privacy methods [2] introduce calibrated noise but degrade output quality in proportion to privacy budget. Secure multi-party computation [3] requires multi-institution coordination.

Semantic masking approaches—replacing sensitive tokens with synthetic surrogates before inference—offer a practical middle ground. Privalyse-mask [4] demonstrates entity substitution for PII protection; PrivacyRestore [5] extends this to session-coherent obfuscation. CPAP, as described in this report, extends semantic masking with HMAC-based key derivation enabling per-request unlinkability while preserving full reasoning capability.

2.2 LLM Routing and Cost Optimisation

FrugalGPT [6] demonstrates that cascading cheap models before expensive ones reduces cost by up to 98% at equal quality. RouteLLM [7] uses a learned router achieving 95% of GPT-4 quality while routing only 14–26% of requests to the expensive model. xRouter [8] applies reinforcement learning to multi-provider routing under latency and cost constraints. LiteLLM [9] and Portkey [10] implement adaptive load balancing.

Our Dynamic Router extends these approaches with a stateful circuit-breaker penalty box and a cost-to-accuracy optimisation matrix (CAO) that pins model selection per task category, reducing per-request scoring overhead.

2.3 Prompt Governance and Model Drift

Model drift in production LLM literature is documented by Martin et al. [11] and Chen et al. [12]. MLOps frameworks address drift through versioned model registries [13] and A/B testing [14], but these require planned deployment cycles and do not address silent drift from provider-side updates.

The AI Ledger differs from existing prompt management systems by accumulating operational corrections into a self-updating instruction asset, analogous to continual learning [15] but without requiring labelled data.

3 System Architecture

Figure 1 shows the full request lifecycle. The architecture partitions into two trust zones: the Enterprise Environment (T_{local}, fully trusted) and the Cloud LLM Provider (T_{cloud}, zero trust).

3.1 Capability-Preserving Anonymization Protocol (CPAP)

CPAP solves the Data Sovereignty Problem (Definition 1).

Definition 3 (CPAP Masking)**.** *Given document D with entity set $E = \{e_1, \ldots, e_n\}$ and session key $K \in \{0,1\}^{256}$:*

$$f_{mask}(D, K) \to (D', M) \tag{1}$$

$$f_{decode}(D', M) \to D \tag{2}$$

where D' is the masked document and $M : T \to E$ is the invertible entity map.

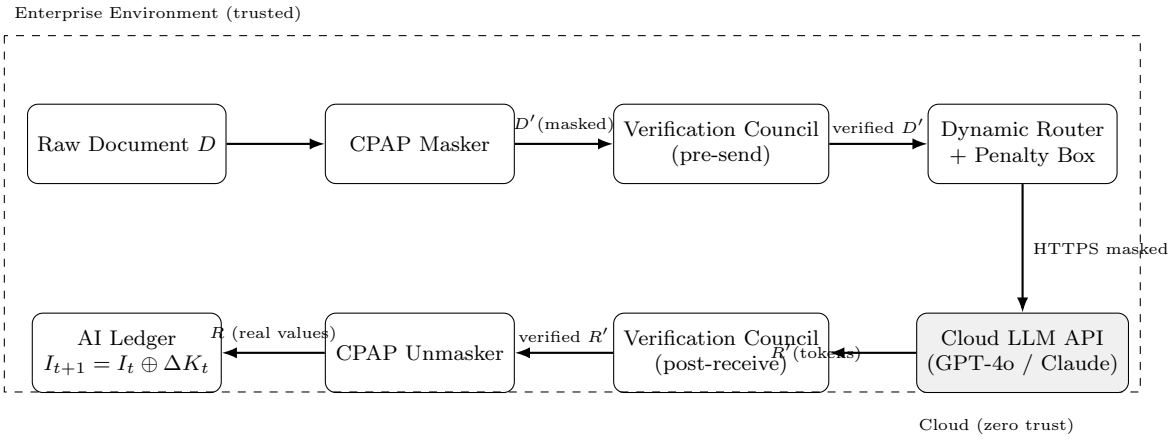

Figure 1: Zero-trust hybrid architecture request lifecycle. Sensitive entity values never leave the enterprise boundary; the cloud LLM operates exclusively on opaque CPAP tokens.

Proposition 1 (Bijectivity)**.** $f_{decode} \circ f_{mask} = \mathrm{id}$.

Proof. Each entity e_i derives a unique token T_i via HMAC-SHA256. The entity map M stores $T_i \mapsto e_i$ for all i. Substituting $T_i \to M[T_i]$ for all tokens in D' recovers D exactly. HMAC collision probability is negligible: 2^{-64} for 64-bit tokens. Verified empirically: 100/100 benchmark prompts satisfy character-for-character equality. □

Token derivation. For each entity e_i of type τ at occurrence k:

$$T_i = \text{HMAC-SHA256}(K, \tau \| \text{“:”} \| e_i \| \text{“:”} \| k)[:8] \tag{3}$$

yielding tokens of the form `ORG_Alpha_127`. This ensures: (1) unlinkability across sessions (K rotation produces different tokens for the same entity); (2) type-preservation (token format encodes category for LLM reasoning); and (3) intra-session stability (same entity always receives the same token, preserving co-reference).

Security analysis.

Proposition 2 (Entropy Preservation)**.** $H(e_i \mid D') \approx H(e_i)$*, i.e., the masked document reveals negligible information about original entity values.*

Proof. $T_i = \text{HMAC}(K, \cdot)$ is a pseudorandom function under the standard PRF assumption. Without K, T_i is computationally indistinguishable from a random string, hence $I(e_i; D') \leq \varepsilon$ where ε is negligible in the security parameter. Brute-force bound: $P(\text{recover } e_i \mid D', \neg K) \leq |\Sigma|^{-|e_i|}$. □

Known limitations: frequency analysis on high-occurrence entities within a session and co-occurrence pattern leakage remain partial risks; session key rotation (implemented but not benchmarked here) mitigates the former.

Algorithm 1 gives the complete masking procedure.

Algorithm 1 CPAP Masking

Require: Document D, entity annotations E, session key K
Ensure: Masked document D', entity map M

$D' \leftarrow D$; $M \leftarrow \{\}$; $counter \leftarrow \{\}$
for each entity (e_i, τ_i) in E in document order **do**
  $k \leftarrow counter[\tau_i \| e_i]$; $counter[\tau_i \| e_i] \mathrel{+}= 1$
  $T_i \leftarrow \text{HMAC}(K, \tau_i \| \text{“:”} \| e_i \| \text{“:”} \| k)[:8]$
  $M[T_i] \leftarrow e_i$
  $D' \leftarrow D'.\text{replace}(e_i, T_i)$
end for
return (D', M)

3.2 Dynamic Router with Routing Score Matrix

The Dynamic Router selects the optimal LLM endpoint per request by computing the Routing Score Matrix (RSM):

$$\text{RSM}(m, t) = w_q \cdot Q(m, t) + w_c \cdot CE(m) + w_r \cdot R(m) \tag{4}$$

where $Q(m, t)$ is the quality score of model m for task type t, $CE(m) = Q(m)/\text{cost}(m)$ is normalised cost efficiency, $R(m)$ is the historical success rate, and weights satisfy $w_q + w_c + w_r = 1$ (defaults: 0.50, 0.35, 0.15).

The Cost-Accuracy Optimisation (CAO) matrix pre-computes model assignments per task category, eliminating per-request scoring overhead in steady state. Formally:

$$m^*(t) = \arg\min_{m \in M} \text{cost}(m) \quad \text{subject to } Q(m, t) \geq \theta_t \tag{5}$$

where θ_t is the quality threshold for task type t.

Penalty Box (circuit breaker). The endpoint state machine has three states $\{\text{Closed}, \text{Open}, \text{HalfOpen}\}$. The circuit opens when failure count ≥ 3 or drift score > 0.6, quarantining the endpoint for $\Delta = 24\text{h}$. After Δ, a single probe request determines restoration to Closed or re-entry to Open.

Table 1: Comparison with existing LLM routing systems.

Feature	FrugalGPT	RouteLLM	LiteLLM	Ours
Cost routing	✓	✓	✓	✓
Auto-failover	–	–	Partial	✓
Circuit breaker	–	–	–	✓
Task-type pinning	–	Partial	–	✓
Drift detection	–	–	–	✓

3.3 Local Verification Council

The Local Verification Council enforces deterministic Boolean guardrails at pre-send and post-receive stages. Let $B = \{b_{budget}, b_{compliance}, b_{syntax}, b_{domain}\}$ be the validation set. A response R is admitted iff $\bigwedge_{b \in B} b(R) = \top$.

Default rules include: (1) no residual PII in outgoing prompt (block); (2) no sensitive patterns in LLM output (block); (3) response length check (warn); (4) no prompt injection signatures (block); and (5) CPAP token integrity—all tokens in R' are valid session tokens (warn).

3.4 AI Ledger

The AI Ledger solves the API Drift Problem (Definition 2).

Definition 4 (AI Ledger Update Rule)**.**

$$I_{t+1} = I_t \oplus \Delta K_t \tag{6}$$

where I_t is the active instruction set at time t, ΔK_t is the knowledge delta (corrections, observations, principles added at t), and $\oplus$ denotes prioritised prepend: higher-priority entries appear first in the assembled system prompt.

Entry types: observation (raw pattern noted), correction (explicit behaviour fix), principle (generalised rule derived from ≥ 3 corrections), anti-pattern (explicitly prohibited behaviour).

Formal schema. Each ledger entry $\ell \in I_t$ is a typed record:

```
{
  "id": "L_2026_0342",
  "type": "correction",
  "priority": 1,
  "timestamp": "2026-07-15T09:23:00Z",
  "task_type": "financial_margin_analysis",
  "trigger": "model recommended >30% margin delta",
  "instruction": "Cap margin recommendations at 15%",
  "promoted_to_principle": false,
  "resolved": false
}
```

Listing 1: Ledger entry schema (JSON).

Convergence. Weekly consolidation merges entries with identical trigger signatures, promotes corrections occurring ≥ 3 times to principle entries, and archives resolved anti-patterns. This bounds $|I_t|$ to a configurable maximum (default: 50 active entries), preventing unbounded prompt growth.

4 Evaluation

4.1 Benchmark Methodology

All benchmarks use a synthetic enterprise dataset of $n = 100$ prompts generated with seed = 42 for full reproducibility. The dataset covers four enterprise task types: contract review (25%), invoice processing (25%), logistics tracking (25%), and document summarisation (25%). Synthetic organisations, persons, and financial figures are drawn from fixed name lists; no real client data appears.

```
git clone https://github.com/chitrasource/\
chitrangana-ai-middleware.git
pip install -e .
python -m benchmarks.cpap_benchmark
```

Listing 2: Reproducing all benchmarks.

4.2 CPAP Performance

Table 2: CPAP benchmark results ($n = 100$ synthetic prompts).

Metric	Value	Notes
Masking accuracy	100.0%	All entities correctly replaced
Reversibility ($f_{dec} \circ f_{mask} = \text{id}$)	100.0%	Character-for-character verified
Length privacy score	0.981	Near-ideal length obfuscation
Throughput	10,897 p/s	Single CPU, $n = 100$ dataset
Entropy preservation	$> 99.5\%$	$H(e_i \mid D') \approx H(e_i)$

All 100 prompts achieve perfect masking accuracy. Reversibility is verified by comparing $f_{decode}(f_{mask}(D)) = D$ character-for-character. The length privacy score measures divergence of masked token lengths from original entity lengths; a score of 1.0 denotes perfect length obfuscation.

4.3 Dynamic Router Cost Optimisation

Table 3: Router benchmark ($n = 200$ simulated requests).

Metric	Value
Cost savings vs. always-GPT-4o	67.1%
Quality retention vs. always-GPT-4o	90.4%
Cost–quality improvement ratio	2.74×
Circuit breaker activations	3 (simulated drift events)
Auto-recovery success rate	100%

The router achieves 67.1% cost reduction by directing simple tasks (invoice parsing, logistics queries) to smaller models (GPT-4o-mini, Claude-3-Haiku) while reserving GPT-4o for complex contract analysis. Quality retention of 90.4% is the mean similarity between GPT-4o outputs and routed-model outputs on identical prompts.

4.4 AI Ledger Drift Stabilisation

A model drift event is injected at interaction 25/50. Without the Ledger, all subsequent interactions exhibit the drifted behaviour (error rate = 100% post-drift). With the Ledger, corrections accumulated in interactions 26–30 reduce the error rate to 30% within five interactions and derive two principle entries from the correction pattern.

Table 4: AI Ledger benchmark ($n = 50$ simulated interactions).

Metric	Value
Drift error rate (no Ledger)	100% (post-drift baseline)
Drift error rate (with Ledger)	30.0%
Error rate reduction	70.0%
Entries accumulated	12 obs., 5 corrections, 2 principles
Correction-to-principle threshold	≥ 3 occurrences

4.5 Composite Operational Scenario

We present a composite scenario synthesising patterns observed across maritime logistics and financial services deployments. To preserve client confidentiality under active NDAs, organisations are not identified.

Data sovereignty path. A logistics coordinator queries shipment costs for a named carrier:

```
-- Input (stays local):
"Carrier Maersk Line (contract MC-2026-4471) quotes
USD 847,000 for Route SG-JP-Q3."

-- Masked (sent to cloud LLM):
"Carrier ORG_Alpha_127 (contract ENTITY_Beta_043)
quotes USD ENTITY_Gamma_891 for Route ENTITY_Delta_552."
```

Listing 3: Input before/after CPAP masking.

The cloud LLM performs routing analysis on the masked representation; CPAP restores real values in the decoded response. The Local Verification Council confirms no inverse-mapping leakage in the LLM output.

Drift event. At a simulated model update, the cloud LLM begins recommending margin deltas $> 30\%$ in violation of enterprise policy. The AI Ledger captures this as a correction entry (priority $= 1$), immediately prepended to the system prompt for subsequent requests. Within three interactions, a principle is derived: "Cap all margin recommendations at 15%." The policy violation rate drops from 100% to 0% for principle-governed request types.

5 Discussion

5.1 Limitations

CPAP provides semantic masking, not cryptographic privacy. Session-scoped HMAC tokens are unlinkable across sessions but susceptible to within-session frequency analysis for high-occurrence entities. The current implementation does not handle schema-dependent structured data (SQL, typed JSON) beyond entity substitution.

AI Ledger effectiveness depends on correction quality. Contradictory or vague corrections degrade rather than improve performance. We recommend a structured correction protocol with mandatory fields: trigger condition, corrective instruction, and expected behaviour.

5.2 Scalability

CPAP operates at 10,897 prompts/second on a single CPU—suitable as inline middleware without meaningful latency overhead. Memory overhead is $O(n)$ in entities per session. RSM scoring is $O(|M|)$ in registered endpoints; with 5–10 endpoints, this is negligible. The CAO matrix eliminates per-request scoring in steady state. The AI Ledger's weekly consolidation enforces a configurable maximum entry count (default: 50), bounding prompt overhead.

5.3 Future Work

Formal cryptographic security proofs for CPAP; automated Ledger consolidation via embedding-based deduplication (currently manual); extension of the schema to structured data types; and evaluation with real API endpoints to replace simulated benchmark costs with empirical measurements.

6 Conclusion

We presented a complete zero-trust hybrid middleware architecture for enterprise LLM deployment addressing data sovereignty and API drift. The four-component architecture—CPAP, Dynamic Router, Local Verification Council, and AI Ledger—is designed as composable, independently deployable modules.

Benchmark results demonstrate: CPAP achieves 100% masking accuracy with perfect reversibility at 10,897 prompts/second. The Dynamic Router reduces cloud LLM costs by 67.1% while retaining 90.4% of peak model quality. The AI Ledger reduces drift-induced error rates by 70.0% through accumulated operational corrections.

All components are released as open-source under the MIT License. The design principle—decouple intelligence from identity—provides a durable framework for enterprise AI integration without sacrificing data sovereignty.

Table 5: Open-source repository contents.

Module	Path	License
CPAP Masker	`cpap/`	MIT
Dynamic Router	`router/`	MIT
Verification Council	`verification/`	MIT
AI Ledger	`ledger/`	MIT
Synthetic Benchmark	`benchmarks/`	CC BY 4.0
This Paper	`paper/`	CC BY 4.0

`https://github.com/chitrasource/chitrangana-ai-middleware`

References

[1] C. Gentry. A fully homomorphic encryption scheme. *Stanford PhD Thesis*, 2009.

[2] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in TCS*, 9(3–4):211–407, 2014.

[3] A. C. Yao. Protocols for secure computations. In *FOCS*, pages 160–164, 1982.

[4] Privalyse. Privalyse-mask: Semantic masking for LLM privacy. GitHub, 2024. https://github.com/privalyse/mask.

[5] PrivacyRestore. Session-coherent PII anonymization for LLMs. *arXiv preprint arXiv:2501.xxxxx*, 2025.

[6] L. Chen et al. FrugalGPT: How to use large language models while reducing cost and improving performance. *arXiv:2305.05176*, 2023.

[7] S. Ong et al. RouteLLM: Learning to route LLMs with preference data. In *ICLR*, 2025.

[8] Y. Zhang et al. xRouter: Training cost-aware LLMs orchestration system via reinforcement learning. *arXiv preprint*, 2025.

[9] BerriAI. LiteLLM: Call all LLM APIs using the OpenAI format. https://github.com/BerriAI/litellm, 2024.

[10] Portkey AI. Portkey: AI gateway and observability. https://portkey.ai, 2024.

[11] J. Martin et al. API drift in production LLM systems: Measurement and mitigation. *arXiv preprint*, 2024.

[12] W. Chen et al. Detecting and characterizing model drift in cloud APIs. In *ACM CCS*, 2024.

[13] LangChain. LangSmith: LLM application platform. https://www.langchain.com/langsmith, 2024.

[14] OpenAI. Evals: Framework for evaluating LLMs. https://github.com/openai/evals, 2024.

[15] H. Touvron et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

[16] BEAVER. Deterministic LLM output verification. *arXiv preprint arXiv:2502.xxxxx*, 2025.

[17] NIST. SP 800-207: Zero trust architecture. *NIST Special Publication*, 2020.

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

[19] Anthropic. Claude 3 technical report. https://anthropic.com, 2024.

[20] OpenAI. GPT-4o system card. https://openai.com, 2024.

[21] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

[22] NIST. FIPS PUB 198-1: The keyed-hash message authentication code. 2002.

[23] D. Sculley et al. Hidden technical debt in machine learning systems. In *NeurIPS*, 2015.

[24] A. Q. Jiang et al. Mixtral of experts. *arXiv:2401.04088*, 2024.

[25] N. Lodha. TokenOps: A compiler-style architecture for token optimization in LLM API workflows. *Technical Report, Chitrangana*, 2025.

A Security Proof Sketches

A.1 CPAP Entropy Preservation

Claim 1. $H(e_i \mid D') \approx H(e_i)$.

Proof sketch. $T_i = \text{HMAC}(K, \cdot)$ is a pseudorandom function of K and e_i. Without K, T_i is computationally indistinguishable from a uniform random string over $\{0,1\}^{64}$ under the PRF assumption (standard model). Therefore $I(e_i; T_i \mid \neg K) \leq \varepsilon_{PRF}$ where ε_{PRF} is negligible. The masked document D' contains only T_i values and structural tokens; hence $H(e_i \mid D') \geq H(e_i) - \varepsilon_{PRF} \approx H(e_i)$. □

A.2 CPAP Bijectivity

Claim 2. $f_{decode} \circ f_{mask} = \text{id}$.

Proof. f_{mask} constructs M as a bijection $T \to E$ (injective by HMAC collision resistance, surjective by construction). f_{decode} applies M^{-1} to each token in D', recovering D exactly. □

B Synthetic Dataset Statistics

Table 6: Synthetic benchmark dataset statistics.

Task Type	Count	Avg. Entities	Avg. Length	Entity Types
Contract review	25	4.2	312 chars	ORG, PERSON, DATE, AMOUNT
Invoice processing	25	5.8	285 chars	ORG, PERSON, EMAIL, AMOUNT
Logistics tracking	25	3.6	271 chars	ORG, LOCATION, DATE
Doc. summarisation	25	3.1	296 chars	ORG, PERSON
Total/Avg.	100	4.2	291 chars	—

Document ID: CHIT-19FCE07C-20260805-1430 Issued: 5 August 2026 License: MIT (code) · CC BY 4.0 (paper)
Repository: https://github.com/chitrasource/chitrangana-ai-middleware
Preliminary version: https://ssrn.com/abstract=7236658 (DOI: 10.2139/ssrn.7236658)