Draft Dokumen Kontrak Pengadaan

1.1 Ruang Lingkup Pekerjaan (Statement of Work / SOW)

STATEMENT OF WORK
Pengadaan Layanan AI Inference API
Proyek GenLaravel

A. Latar Belakang

Proyek GenLaravel memerlukan layanan AI inference untuk menjalankan 10 AI agents yang berfungsi menghasilkan Laravel UI components dari natural language prompts. Layanan ini harus mampu memproses request dengan kecepatan tinggi, akurasi output yang baik, dan ketersediaan yang reliable.

B. Tujuan Pengadaan

Pengadaan layanan AI inference API bertujuan untuk:

1. Menyediakan infrastruktur AI inference yang reliable dan scalable
2. Mendukung operasional 10 AI agents dalam pipeline generation
3. Memastikan response time yang cepat untuk user experience yang optimal
4. Meminimalkan biaya operasional dengan model pay-as-you-go

C. Ruang Lingkup Layanan

Vendor menyediakan layanan sebagai berikut:

1. **AI Model Access**

• Akses ke Large Language Model untuk text generation

• Model yang mendukung coding tasks dan natural language understanding

• Kemampuan untuk memproses prompt kompleks (multi-turn conversation)

2. **API Infrastructure**

• RESTful API endpoint yang secure (HTTPS)

• Authentication menggunakan API key

• Rate limiting yang reasonable untuk development dan production

3. **Performance**

• Inference speed minimal 1000 tokens/second

• Latency maksimal 500ms untuk first token

• Uptime minimal 99.5%

4. **Support**

- Dokumentasi API yang lengkap
- Status page untuk monitoring availability
- Support channel untuk technical issues

D. Deliverables

No	Deliverable	Deskripsi	Timeline
1	API Access	API key dan endpoint access	Day 1
2	Documentation	API documentation dan integration guide	Day 1
3	Usage Dashboard	Dashboard untuk monitoring usage dan billing	Day 1
4	Support Access	Akses ke support channel	Day 1

E. Periode Kontrak

Kontrak berlaku selama periode proyek (14 minggu) dengan opsi perpanjangan berdasarkan kebutuhan. Model pricing pay-as-you-go memungkinkan fleksibilitas tanpa commitment jangka panjang.

F. Lokasi Pekerjaan

Layanan disediakan secara cloud-based, dapat diakses dari lokasi manapun dengan koneksi internet.

G. Asumsi dan Batasan

Asumsi:

- Koneksi internet tersedia dan stabil
- Tim memiliki kemampuan untuk integrasi API
- Usage tidak melebihi rate limit yang ditetapkan

Batasan:

- Layanan terbatas pada API inference, tidak termasuk fine-tuning
- Data yang dikirim ke API tidak disimpan permanen oleh vendor
- Tidak ada dedicated support, menggunakan community/standard support

1.2 Service Level Agreement (SLA)

SERVICE LEVEL AGREEMENT
Layanan AI Inference API

A. Definisi Layanan

Layanan AI Inference API mencakup akses ke endpoint API untuk melakukan text generation menggunakan Large Language Model yang disediakan vendor.

B. Availability

Metric	Target	Measurement
Monthly Uptime	99.5%	(Total Minutes - Downtime Minutes) / Total Minutes × 100%
Scheduled Maintenance	Maksimal 4 jam/bulan	Dengan notifikasi 24 jam sebelumnya
Unscheduled Downtime	Maksimal 2 jam/incident	Measured from incident report to resolution

C. Performance Metrics

Metric	Target	Measurement
Inference Speed	Minimal 1000 tokens/second	Average across all requests
Time to First Token	Maksimal 500ms	P95 latency
Request Success Rate	99%	Successful requests / Total requests
Error Rate	Maksimal 1%	Failed requests / Total requests

D. Support Response Time

Priority	Description	Response Time	Resolution Time
Critical	Service completely unavailable	1 jam	4 jam
High	Significant performance degradation	4 jam	24 jam
Medium	Minor issues, workaround available	24 jam	72 jam
Low	Questions, feature requests	72 jam	Best effort

E. Kompensasi (Service Credits)

Jika SLA tidak terpenuhi, kompensasi diberikan dalam bentuk service credits:

Monthly Uptime	Service Credit
99.0% - 99.5%	10% of monthly bill

95.0% - 99.0%	25% of monthly bill
< 95.0%	50% of monthly bill

F. Exclusions

SLA tidak berlaku untuk:

- Downtime akibat scheduled maintenance yang sudah dinotifikasi
- Issues yang disebabkan oleh client-side (network, integration bugs)
- Force majeure (bencana alam, perang, dll)
- Penggunaan yang melanggar terms of service

G. Monitoring dan Reporting

- Vendor menyediakan status page yang dapat diakses publik
- Usage dan billing dapat dimonitor melalui dashboard
- Monthly report tersedia untuk enterprise customers

1.3 RACI Matrix (Vendor vs Organisasi)

RACI MATRIX

Proyek GenLaravel - Vendor Management

Keterangan:

- **R** = Responsible (melaksanakan pekerjaan)
- **A** = Accountable (bertanggung jawab atas hasil)
- **C** = Consulted (dimintai pendapat)
- **I** = Informed (diberi informasi)

Aktivitas	**Project Manager (Fikri)**	**Frontend Dev (Nadia)**	**Cerebras (Vendor)**	**Mistral (Vendor)**	**Security Consultant**
API Integration					
API key management	A, R	I	C	C	I
Integration development	A	R	C	C	I
Error handling implementation	A	R	C	C	I

Fallback mechanism setup	A, R	C	I	I	I
Operations					
Usage monitoring	A, R	I	R	R	I
Cost tracking	A, R	I	R	R	I
Performance monitoring	A, R	C	R	R	I
Incident response	A	C	R	R	C
Security					
API key security	A, R	C	I	I	C
Data handling compliance	A	I	R	R	C
Security audit	A	I	I	I	R
Vulnerability assessment	A	C	C	C	R
Support					
Technical support requests	R	R	A	A	I
Escalation management	A, R	I	R	R	I
Documentation updates	A	R	R	R	I
Contract Management					
Contract negotiation	A, R	I	R	R	I
SLA monitoring	A, R	I	R	R	I
Invoice processing	A, R	I	R	R	I
Contract renewal	A, R	I	C	C	I
Development					
Agent development	A, R	C	I	I	I
Frontend development	A	A, R	I	I	I
Testing	A	R	C	C	C

Deployment	A, R	C	I	I	C

1.4 Alokasi Risiko dalam Kontrak

RISK ALLOCATION MATRIX

No	Risiko	Deskripsi	Probabilitas	Dampak	Penanggung Risiko	Mitigasi
1	API Downtime	Layanan API tidak tersedia	Rendah	Tinggi	Vendor (Cerebras/Mistral)	SLA dengan service credits, fallback ke vendor alternatif
2	Rate Limiting	Request dibatasi karena melebihi quota	Sedang	Sedang	Organisasi	Implementasi request queuing, batch processing, monitoring usage
3	Price Increase	Vendor menaikkan harga API	Rendah	Sedang	Organisasi	Kontrak dengan price lock period, evaluasi vendor alternatif
4	Data Breach	Data sensitif bocor melalui API	Rendah	Tinggi	Shared (Vendor 60%, Organisasi 40%)	Tidak mengirim data sensitif, API key rotation, audit trail
5	Model Deprecation	Model yang digunakan di-deprecate	Sedang	Sedang	Vendor	Notifikasi 90 hari sebelum deprecation, migration support
6	Integration Failure	Integrasi API gagal atau bermasalah	Sedang	Tinggi	Organisasi	Thorough testing, error handling,

						rollback mechanism
7	Vendor Lock-in	Ketergantungan tinggi pada satu vendor	Sedang	Sedang	Organisasi	Multi-vendor strategy (Cerebras + Mistral), abstraction layer
8	Compliance Issues	Pelanggaran terms of service	Rendah	Tinggi	Organisasi	Review ToS secara berkala, compliance training
9	Performance Degradation	Response time meningkat signifikan	Sedang	Sedang	Vendor	SLA dengan performance metrics, monitoring alerts
10	Support Unavailability	Support tidak responsif	Rendah	Sedang	Vendor	SLA dengan response time commitment, escalation path

Pembagian Tanggung Jawab Risiko:

Kategori Risiko	**Vendor**	**Organisasi**
Infrastructure & Availability	80%	20%
Security & Compliance	60%	40%
Integration & Development	20%	80%
Cost Management	30%	70%
Performance	70%	30%

1.5 Acceptance Criteria

ACCEPTANCE CRITERIA
Layanan AI Inference API

A. Kriteria Penerimaan Teknis

No	Kriteria	Metode Verifikasi	Target	Status
1	API endpoint accessible	HTTP request test	Response 200 OK	Pass/Fail
2	Authentication working	API key validation	Successful auth	Pass/Fail
3	Inference speed	Benchmark test (100 requests)	≥ 1000 tokens/s average	Pass/Fail
4	Time to first token	Latency measurement	≤ 500ms (P95)	Pass/Fail
5	Output quality	Sample prompt testing	Coherent, relevant output	Pass/Fail
6	Error handling	Error scenario testing	Proper error codes & messages	Pass/Fail
7	Rate limit compliance	Load testing	Documented limits enforced	Pass/Fail
8	Documentation accuracy	Manual verification	Docs match actual behavior	Pass/Fail

B. Kriteria Penerimaan Fungsional

No	Kriteria	Deskripsi	Acceptance Test
1	Code Generation	API dapat menghasilkan kode Laravel/Blade yang valid	Generate 10 sample components, verify syntax
2	Natural Language Understanding	API memahami prompt dalam bahasa Indonesia dan Inggris	Test dengan 20 prompts berbeda
3	Context Handling	API dapat memproses multi-turn conversation	Test conversation dengan 5+ turns
4	Long Output	API dapat menghasilkan output panjang (>2000 tokens)	Generate full page template
5	Consistency	Output konsisten untuk prompt yang sama	Test 5x dengan prompt identical

C. Kriteria Penerimaan Operasional

No	Kriteria	Deskripsi	Verification
1	Dashboard Access	Dapat mengakses usage dashboard	Login dan view usage
2	Billing Transparency	Billing jelas dan sesuai usage	Compare usage vs invoice

3	Support Responsiveness	Support merespons dalam SLA	Submit test ticket
4	Status Page	Status page accessible dan akurat	Monitor selama 1 minggu

D. Proses Acceptance Testing

1. **Preparation Phase (Day 1-2)**

• Setup test environment

• Prepare test cases dan test data

• Configure monitoring tools

2. **Execution Phase (Day 3-5)**

• Execute technical acceptance tests

• Execute functional acceptance tests

• Document results

3. **Evaluation Phase (Day 6-7)**

• Review test results

• Identify issues dan gaps

• Vendor remediation jika diperlukan

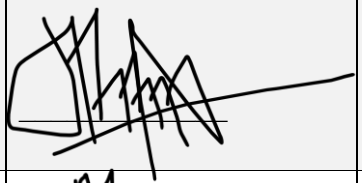
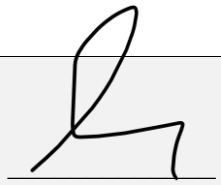
4. **Sign-off Phase (Day 8)**

• Final review

• Acceptance sign-off atau rejection dengan justifikasi

• Contract activation

E. Acceptance Sign-off

Role	Name	Signature	Date
Project Manager	Fikri Armia Fahmi		//2025
Technical Lead	Fikri Armia Fahmi		//2025
Vendor Representative			//2025

F. Rejection Criteria

Layanan akan ditolak jika:

- Lebih dari 2 kriteria teknis tidak terpenuhi
- Kriteria fungsional nomor 1 (Code Generation) tidak terpenuhi
- Uptime selama testing period < 95%
- Response time rata-rata > 1 detik
- Vendor tidak dapat menyediakan dokumentasi yang memadai

8. Kesimpulan

Analisis Make-or-Buy untuk proyek GenLaravel menghasilkan keputusan strategis yang mengoptimalkan penggunaan sumber daya. Dengan memilih MAKE untuk core competency (AI agents, frontend UI, Laravel integration) dan BUY untuk komponen pendukung (AI inference API, hosting, tools), proyek dapat:

1. Fokus pada pengembangan nilai tambah utama (AI agents dan queue system)
2. Menghemat biaya signifikan (Rp 65-120 juta untuk AI infrastructure)
3. Mempercepat time-to-market dengan memanfaatkan solusi existing
4. Mempertahankan intellectual property untuk komponen strategis
5. Mengurangi risiko teknis dengan menggunakan vendor yang proven
6. Menjaga simplicity dengan asyncio.Queue untuk MVP, dengan path upgrade yang jelas ke distributed queue jika diperlukan

Draft dokumen kontrak pengadaan yang disusun mencakup SOW, SLA, RACI Matrix, alokasi risiko, dan acceptance criteria yang komprehensif untuk memastikan hubungan vendor yang profesional dan terkelola dengan baik.