

LLM Booster SDK Developer Edition

Kurumlar için değerlendirme, dağıtım ve teknik inceleme kılavuzu

Lisans kapsamı: Yalnızca kurum içi değerlendirme, test ve geliştirme. Üretim/ticari kullanım ve üçüncü taraflara dağıtım bu sürümde lisanslı değildir. Özellikler teknik olarak kapatılmaz. Tam koşullar paket içindeki LICENSE-DEVELOPER.md belgesindedir.

1. Amaç ve sürüm kapsamı

Bu belge, kurumların LLM Booster Developer Edition paketini kontrollü bir teknik değerlendirme ortamında incelemesine yardımcı olur. Sürüm; ajan bağlamı hazırlama, anahtar sözcük tabanlı ajan seçimi, metin temizleme ve deneysel taşıma kodlaması işlevleri sunan Python SDK'sıdır. Python 3.10 veya üzeri gerekir; çalışma zamanı bağımlılığı tanımlanmamıştır.

Developer Edition ticari/kurumsal üretim lisansı değildir; SLA, kurumsal destek, güvenlik sertifikası, güncelleme taahhüdü veya uyumluluk onayı sağlamaz. Kurum içinde değerlendirmeye katılacak yetkili çalışan ve yükleniciler, lisans metnindeki erişim koşullarına uymalıdır.

2. Paket ve kurulum

Dağıtım ZIP'i şu bileşenleri taşır: Python wheel paketi, LICENSE-DEVELOPER.md, SDK Kullanım Kılavuzu PDF'si ve bu Kurum Kılavuzu PDF'si. Wheel çevrimdışı kurulum için kullanılabilir; Python sürümü ve mimari hedefi paketin metadata'sına göre doğrulanmalıdır. Python 3.10 veya üzeri kurulu olmalıdır. PowerShell komutlarını ZIP'in indirildiği klasörde çalıştırın:

```
py -0p
py -3.10 --version
Expand-Archive .\llm_booster_developer_edition-0.1.0.zip .\llm-booster-dev
Set-Location .\llm-booster-dev
py -3.10 -m venv .venv
$wheel = Get-ChildItem -Filter 'llm_booster-*.whl' | Select-Object -First 1
if (-not $wheel) { throw "Developer Edition wheel not found in the extracted folder." }
.\.venv\Scripts\python.exe -m pip install $wheel.FullName
.\.venv\Scripts\python.exe -m pip show llm-booster
.\.venv\Scripts\llm-booster.exe --help
```

Kaynak depodan geliştirme kurulumu

ZIP yerine SDK kaynak deposuyla çalışıyorsanız aşağıdaki komutları depo kökünde çalıştırın:

```
py -3.10 -m venv .venv
.\.venv\Scripts\python.exe -m pip install -e .
```

Sanal ortam, kurulumu sistem Python'undan ayırır. ZIP başka bir klasöre açıldıysa Expand-Archive komutundaki arşiv yolunu ve hedef klasörü değiştirin. Kurum paketleri kurum içi artifact deposunda tutulabilir; dış kuruluşlara veya müşterilere aktaramaz. Yeniden dağıtım izni ancak ayrı yazılı sözleşmeyle verilebilir.

2.1. CLI ile örnek çalışma

Aşağıdaki örnek, test bağlamını UTF-8 JSON dosyasına yazar ve CLI ile temizleme/yönlendirme kiplerini çalıştırır. CLI çıktısı JSON'dur; bu komutlar bir LLM servisine istek göndermez.

```
@'
{
  "billing_agent": "Invoice 426 TL is unpaid.",
  "device_agent": "SIM 123 is active.",
  "support_agent": "Customer requested escalation."
}
'@ | Set-Content -Encoding utf8 .\context.json
```

```
$cli = ".\venv\Scripts\llm-booster.exe"
& $cli run --mode baseline --request invoice --context .\context.json
& $cli run --mode cleanup --request invoice --context .\context.json
& $cli run --mode semantic --top-k 2 --request invoice --context .\context.json
& $cli run --mode linear_network_coding --request invoice --context .\context.json
```

3. Teknik çalışma ve veri akışı

Bileşen	SDK davranışı	Kurum incelemesi
Booster.run	İstek ve ajana ait bağlamı yerel olarak işler; bu çağrı LLM servisi çağırır.	İçeriği uygulamanız ve süreç belleği içinde yönetin.
semantic modu	Ajanları metin anahtar sözcüklerinin örtüşmesine göre sıralar.	Model tabanlı anlamsal anlama veya kalite garantisi değildir.
call_local_ollama	Açıkça çağrılırsa varsayılan olarak localhost:11434 üzerindeki Ollama API'sine HTTP isteği yapar.	Model, ağ, log ve veri saklama ayarlarını kurum ayrıca denetler.
Kodlama yardımcıları	GF(2) tek paket silinmesini karşılayabilir; GF(256) roundtrip ayrı deney API'sidir.	Şifreleme, kimlik doğrulama veya bozulma tespiti sağlamaz.

Veri sorumluluğu: SDK'ya verilen iş verileri, kişisel veriler, gizli bilgiler ve müşteri verileri kurumun sınıflandırma ve kullanım politikalarına tabidir. Bu sürüm için telemetri, saklama, şifreleme veya mevzuat uyumu hakkında bağımsız bir taahhüt verilmez. Hassas verilerle değerlendirme yapmadan önce kurum güvenlik ve hukuk ekiplerinin onayını alın.

3.1. Ollama ile yerel modeller

SDK'nin varsayılan Booster.run ve CLI akışları yalnızca yerel bağlam hazırlığı yapar; model çalıştırmaz. Ollama'nın Windows uygulamasını kurup başlatın (veya servis çalışmıyorsa ayrı bir PowerShell'de ollama serve komutunu çalıştırın). İlk model indirmeleri internet ve disk alanı gerektirir; indirilen modellerle çıkarım Ollama'nın yerel HTTP API'si üzerinden yapılır.

Qwen 2.5 3B ve NVIDIA Nemotron Mini 4B

```
ollama --version
ollama pull qwen2.5:3b
ollama run qwen2.5:3b
# Ollama isteminden çıkmak için: /bye
ollama pull nemotron-mini:4b
ollama run nemotron-mini:4b
```

Modeli ollama run ile doğruladıktan sonra aynı yerel modele llm_booster.call_local_ollama ile istek gönderin. Model adını Ollama'da kurulu başka bir model etiketiyle değiştirebilirsiniz:

```
@'
from llm_booster import call_local_ollama

model = "qwen2.5:3b" # Diğer seçenekler: nemotron-mini:4b veya kumru:2b.
prompt = "Türkçe yanıt ver: Fatura 426 TL için neyi kontrol etmeliyim?"
result = call_local_ollama(prompt, model=model)
if not result["ok"]:
    raise RuntimeError(result["error"])
print(result["text"])
'@ | Set-Content -Encoding utf8 .\local_model_demo.py
.\.venv\Scripts\python.exe .\local_model_demo.py
```

Kumru 2B GGUF modelini Ollama'ya ekleme

Kumru örneğinde Hugging Face'teki icecubetr/Kumru-2B-GGUF deposundan Q4_K_M GGUF dosyası indirilir ve Ollama Modelfile ile yerel bir model etiketi oluşturulur. İndirme aracı SDK'nın zorunlu bağımlılığı değildir:

```
.\.venv\Scripts\python.exe -m pip install huggingface_hub
New-Item -ItemType Directory -Force .\models\kumru | Out-Null
.\.venv\Scripts\hf.exe download `
    icecubetr/Kumru-2B-GGUF kumru-2b-Q4_K_M.gguf `
    --local-dir .\models\kumru
'@'
FROM ./models/kumru/kumru-2b-Q4_K_M.gguf
'@ | Set-Content -Encoding ascii .\Modelfile
ollama create kumru:2b -f .\Modelfile
ollama run kumru:2b
```

Model seçimi ve lisans: Etiketler ve model kartları değişebilir; model deposundaki lisansı, kullanım koşullarını ve indirdiğiniz GGUF dosyasının kaynağını kurum politikasına göre doğrulayın. Örneğin Qwen2.5 3B'nin lisansı diğer Qwen2.5 boyutlarından farklıdır; Kumru GGUF deposu temel modelin Apache 2.0 lisansını belirtir. Nemotron Mini 4B İngilizce kullanım için optimize edilmiştir; Türkçe sonuç kalitesini kurumunuz ayrıca değerlendirmelidir. Model ağırlıklarının indirilmesi ücretsiz olsa da donanım maliyeti olabilir ve bu SDK'nın Developer Edition lisansı üretim/ticari kullanım izni vermez.

SDK'de Ollama varsayılanı <http://localhost:11434>'tür. call_local_ollama, çağrıda hata olursa ok: false ve error alanlarını döndürür. Ajan başına yerel model çağrısı örneği için SemanticRouter().route_local_model(..., model="qwen2.5:3b") kullanılabilir; bu işlev de seçilen her ajan için Ollama API'sine ayrı bir istek yapar.

4. İzin verilen kullanım ve sınırlar

İzin verilen	Bu lisansla izin verilmeyen
Kurum içi sandbox ortamında değerlendirme, test ve geliştirme.	Üretim iş yükleri veya müşteriye/son kullanıcıya sunulan ticari hizmet.
Lisans metnine bağlı kurum çalışanları/yüklenicilerinin kontrollü erişimi.	Paketin müşterilere, bağlı olmayan iştiraklere veya diğer üçüncü taraflara aktarılması.
Lisans kapsamı için gerekli, kontrollü kurum içi kopyalar.	Yeniden satış, alt lisans, OEM/embedded ürün veya kamuya açık paket deposu.
Kod ve performansın kurum içi teknik incelemesi.	SDK'yı sertifikalı, SLA kapsamlı veya üretime hazır olarak tanıtmak.

5. Güvenlik ve tedarik kontrol listesi

- Lisans dosyasını hukuk/tedarik birimine inceletin; Lisans Veren kimliği ve sözleşme kabul sürecini doğrulayın.
- Wheel SHA-256 özetini kurum içi kayıtlı saklayın ve yüklenen paketin kaynağını doğrulayın.
- Yalnızca izole geliştirme ortamında ve hassas olmayan sentetik veriyle başlayın.
- Dış modele yönlendirme yapan uygulama kodu veya call_local_ollama kullanımını ayrı ağ ve veri akışı incelemesine tabi tutun.
- Günlük, hata raporu ve test çıktılarında gizli/kişisel veri bulunmadığını kontrol edin.
- Üretim ihtiyacı oluşursa Developer Edition'ı üretime taşımayın; ayrı ticari/enterprise lisans ve teknik değerlendirme talep edin.

6. Doğrulama

```
.\.venv\Scripts\python.exe -m pip show llm-booster
.\.venv\Scripts\python.exe -m unittest discover -s tests -v
```

Test komutu kaynak kod paketinin yanında sağlanan test dizini mevcutsa çalışır. Yalnızca wheel alan kurumlar için kabul testi, CLI yardımını ve kurumdaki örnek bağlamla test çağrısını doğrulamalıdır.

7. Ticari kullanım veya yaygın dağıtım ihtiyacı

Developer Edition bu ihtiyaçları karşılamaz. Üretim, ticari servis, kurum genelinde/bağlı şirketlerde yaygın kullanım, SLA, destek, OEM veya ürüne gömme hakları ayrı bir sözleşme ve lisans kapsamı gerektirir. Bu belgede fiyat, teslim süresi veya destek seviyesi taahhüt edilmemiştir.

Yasal durum notu

LICENSE-DEVELOPER.md dağıtım öncesi hukuk incelemesi gereken taslaktır; Lisans Veren tüzel kişi, geçerli hukuk, yargı yeri ve bildirim bilgileri henüz tanımlı değilse kuruma gönderilmeden önce tamamlanmalıdır. Bu kılavuz güvenlik denetimi veya hukuki görüş yerine geçmez.

Doküman sürümü: Developer Edition 0.1.0 • Kaynak depo: LLM-BOOSTER-SDK