Leistungsbeschreibung & Preisindikation

GPU-/Compute-Infrastruktur für privates KI-Inferencing (Forschung & Entwicklung)

FFG Basisprogramm – Antrag Nr. 71545616 · Antragsteller: Plappi FlexKapG (in Gründung) · Kostenposition: Sachkosten „GPU/Compute (privates Inferencing, F&E)“, ~90.000 EUR (Forschungsjahr 1)

1. Anbieter

Scaleway SAS – europäischer Cloud-Anbieter mit Rechenzentren in der EU (Frankreich/Niederlande). DSGVO-konform, EU-Datenhaltung. Öffentliche Preisliste: scaleway.com/en/pricing.

2. Leistungsumfang

Nutzungsbasierte Anmietung von GPU-Recheninstanzen (on-demand, stundenweise Abrechnung) für Training/Finetuning mehrsprachiger Sprach- und Dialogmodelle sowie für den Aufbau und die Erprobung der privaten EU-Inferenz. Die Abrechnung erfolgt verbrauchsabhängig (nach Bedarf abgerufen), nicht als Vorab-Festkauf.

3. Warum kein klassisches Festangebot

Die Wahl der konkreten GPU-Modelle (z. B. H100, L40S, L4, A100) ist Gegenstand der Forschung: Welche GPU je Modell/Workload das beste Verhältnis aus Qualität, Latenz und Kosten liefert, wird im Projekt iterativ evaluiert und optimiert. Bei on-demand-Cloud-Compute ist daher ein fixes Einzelangebot nicht vorgesehen; Beschaffungsgrundlage ist die öffentliche Stunden-Preisliste des Anbieters.

4. Indikative Listenpreise (Stand 2026, € je GPU-Stunde)

GPU-Instanz	Speicher	~ € / Stunde (indikativ)
H100	80 GB	~ 2,73
L40S	48 GB	~ 1,40
L4	24 GB	~ 0,75
A100-Klasse / RENDER-S	40–80 GB	~ 1,00 – 2,00

5. Budgetierung

Geplantes Volumen ~90.000 EUR über Forschungsjahr 1, verbrauchsabhängig abgerufen (Mischung aus Trainings-/Finetuning-Läufen und Inferenz-Erprobung). Bei einem Mischsatz von ~1,5 €/GPU-Stunde entspricht dies grob ~60.000 GPU-Stunden bzw. mehreren parallel betriebenen Instanzen über die Laufzeit.

Hinweis: Diese Leistungsbeschreibung dokumentiert Anbieter, Leistungsumfang und Preisgrundlage gemäß FFG-Anforderung für Kostenpositionen über 20.000 EUR. Eine verbindliche Abrechnung erfolgt nutzungsbasiert nach tatsächlichem Verbrauch.