# Documentație Tehnică: Strategia de Utilizare a Modelelor AI

**Obiectiv:** Definirea unei strategii clare pentru selectarea modelului AI adecvat pentru fiecare sarcină din cadrul platformei "AI Play & Learn Studio", pentru a optimiza costurile, viteza și calitatea rezultatelor.

### **1. Principiul de Bază: Rutare Inteligentă**

Asistentul nostru, Sparky, va acționa ca un dispecer inteligent. În loc să folosească un singur model pentru toate cererile, motorul nostru de rutare a intențiilor (sparky-intent-router) va analiza natura sarcinii și va invoca modelul cel mai potrivit.

### **2. Alocarea Sarcinilor pe Modele**

Vom împărți sarcinile creative și administrative pe diferite modele pentru a maximiza eficiența.

#### **Utilizarea Gemini 2.5 Pro (Modelul de Gândire Complexă)**

Acest model va fi rezervat pentru sarcinile care necesită cea mai mare creativitate, coerență și înțelegere a contextului.

* **Sarcină Principală: Generarea Conținutului de Bază al Atelierelor**
  + **Generarea Poveștilor:** Când un utilizator cere "o poveste despre un pui de dinozaur timid", vom folosi **Pro** pentru a asigura o narațiune de înaltă calitate, cu personaje bine conturate.
  + **Crearea Ghidului Pedagogic:** La generarea pachetului final, textul pentru ghidul părintelui, care necesită structură și explicații clare, va fi generat de **Pro**.

**Justificare:** Acestea sunt sarcinile cu cea mai mare valoare percepută de client. Calitatea lor este non-negociabilă.

#### **Utilizarea Gemini 2.5 Flash / Flash-Lite (Modelele Rapide și Eficiente)**

Aceste modele vor fi folosite pentru sarcini de volum mare, rapide, unde viteza de răspuns este crucială și complexitatea creativă este mai mică.

* **Sarcini Principale: Asistență și Optimizare**
  + **"Promptul Magic" din Dashboard:** Când un utilizator scrie *"un atelier despre animale polare cu un joc"*, vom folosi **Flash-Lite** pentru a extrage rapid entitățile (tema: animale polare, componente: joc) și a pre-completa blueprint-ul.
  + **Generarea de Metadate:**
    - Sugerarea de titluri SEO pentru listările Etsy/TPT.
    - Generarea celor 13 tag-uri relevante.
    - Crearea descrierilor scurte pentru proiecte.
  + **Funcționalități de Chat și Suport:** Răspunsurile la întrebările simple din viitoarea fereastră de chat cu Sparky.

**Justificare:** Aceste sarcini sunt frecvente și necesită un răspuns aproape instantaneu pentru a menține fluiditatea experienței utilizatorului. Costul redus per apel ne permite să oferim aceste funcționalități la scară largă.

#### **Utilizarea Modelelor Specializate (Imagen, Veo etc.)**

Acestea sunt invocate atunci când intenția este clară.

* **Imagen:** Pentru orice cerere de generare de imagine statică.
* **Veo / Gemini Speech:** Pentru funcționalitățile din **Faza 5 (Expansiunea Multimedia)**.

### **3. Implementarea în Arhitectură**

Funcția noastră Cloud sparky-intent-router va fi actualizată. Pe lângă identificarea intenției (ex: generate\_story, suggest\_tags), va avea și o logică de selecție a modelului.

**Pseudo-cod pentru router:**

def intent\_router(request):  
 user\_prompt = request.json['prompt']  
 intent = analyze\_intent(user\_prompt)  
  
 if intent == 'generate\_full\_story':  
 # Apel către funcția care folosește Gemini Pro  
 return call\_pro\_model\_service(user\_prompt)  
 elif intent == 'suggest\_etsy\_tags':  
 # Apel către funcția care folosește Gemini Flash-Lite  
 return call\_flash\_lite\_service(user\_prompt)  
 elif intent == 'generate\_image':  
 # Apel către funcția care folosește Imagen  
 return call\_imagen\_service(user\_prompt)  
 # ... și așa mai departe

Prin adoptarea acestei strategii, construim un produs care nu este doar puternic, ci și **inteligent din punct de vedere economic și tehnic**. Oferim cea mai înaltă calitate acolo unde contează cel mai mult și cea mai mare viteză pentru sarcinile de zi cu zi.