Stories We Tell - LLM Backend Development Proposal

Phases 1-4: Remaining Work

Date: December 2025
Client Focus: LLM Backend Only (Phases 1-4)
Excluded: Phase 5-6 (Visual Generation & Final Assembly - handled by external VLM teams)

Executive Summary

This proposal outlines the **remaining work** needed to complete the LLM-powered story intake and narrative production pipeline. Most of the system is already built and operational.

What's Already Complete:

- ✅ Phase 1: Intake Chatbot + Story Record Engine (80% complete)
- ✅ Phase 2: Admin Review Dashboard (100% complete)
- ✅ Phase 3: Script Generation (90% complete - script generation works)
- ✅ Phase 4: Script Review Interface (100% complete)

What Remains:

- Phase 1: Enhanced metadata collection + RAG indexing (5-7 days)
- Phase 3: Synopsis generation only (2-3 days)

Total Remaining Work: 7-10 days
Monthly Cost: ~$19-32/month (LLM API usage - client provides API keys)

Current Status

✅ Already Built & Operational

Phase 1 (80% Complete)

- ✅ Conversational intake chatbot
- ✅ Real-time AI streaming responses
- ✅ Story dossier extraction (characters, scenes, locations, plot)
- ✅ Photo/document uploads
- ✅ Multi-user authentication
- ✅ Session management
- ✅ Project-based organization

Phase 2 (100% Complete)

- ✅ Admin dashboard (`/admin`)
- ✅ Validation queue system
- ✅ **Full conversation transcript viewing**
- ✅ **Script viewing and editing**
- ✅ Approve → sends email to client (with CC to business emails)
- ✅ Reject with feedback notes

Phase 3 (90% Complete)

- ✅ Script generation (automatically generates when story completes)
- ✅ Script stored in validation queue
- ✅ Uses Claude Sonnet 4.5 or GPT-4.1
- ✅ 3-5 minute video script format

Phase 4 (100% Complete)

- ✅ Script review interface
- ✅ Script editing before approval
- ✅ Email delivery on approval

Remaining Work

Phase 1 Completion: Enhanced Metadata + RAG Indexing (5-7 days)

1.1 Enhanced Metadata Collection (2-3 days)

What to Build:

- Extend dossier extraction to capture:
  - `era`: Timeline/period (e.g., "1940s", "Medieval", "Contemporary")
  - `audience`: Target audience (e.g., "Family", "Adult", "Teen")
  - `emotional_perspective`: Emotional tone (e.g., "Nostalgic", "Tense", "Hopeful")
  - `environmental_metadata`: Weather, lighting, atmosphere

Changes:

- Update `app/ai/dossier_extractor.py` to extract new fields
- Update `dossier.snapshot_json` schema
- Migration: `20250118000001_add_enhanced_metadata.sql`

APIs: No new endpoints needed (uses existing `/api/v1/chat` and `/api/v1/dossier`)

Cost: ~$0.01-0.02 per story (GPT-4o extraction) = ~$1-2/month (100 stories)

1.2 Story Record Indexing & RAG Layer (3-4 days)

What to Build:

- Story embedding generation (OpenAI `text-embedding-3-small`)
- Vector storage in Supabase (pgvector extension)
- Similarity search for related stories

Database Changes:

```
CREATE TABLE story_embeddings (
    story_id UUID PRIMARY KEY REFERENCES dossier(project_id),
    embedding vector(1536),
    created_at TIMESTAMPTZ DEFAULT NOW()
);
CREATE INDEX ON story_embeddings USING ivfflat (embedding
vector_cosine_ops);
```

APIs:

- `POST /api/v1/stories/{story_id}/index` - Index a story
- `GET /api/v1/stories/{story_id}/similar` - Find similar stories
- `GET /api/v1/stories/search?query=...` - Semantic search

Cost: ~$0.0001 per story (embeddings) = ~$0.01/month (negligible)

Phase 3: Synopsis Generation (2-3 days)

What to Build:

- Add synopsis generation to existing script generation flow
- Generate synopsis from dossier data before/alongside script
- Store synopsis in validation queue

Implementation:

- Add `_generate_synopsis_response` method to `app/ai/models.py`
- Add `TaskType.SYNOPSIS` enum value
- Modify `app/api/simple_chat.py` story completion flow
- Add `synopsis` field to `validation_queue` table

Database:

```
ALTER TABLE validation_queue
ADD COLUMN IF NOT EXISTS synopsis TEXT;
```

APIs: No new endpoints (synopsis generated automatically, stored in validation queue)

Cost: ~$0.03-0.05 per synopsis (GPT-4o-mini or GPT-4o) = ~$3-5/month (100 stories)

Cost Breakdown

LLM API Costs (Client provides API keys)

Note: All costs are estimates using current API pricing. Actual costs may vary based on conversation length and model usage.

Service	Usage	Cost per Story	Monthly (100 stories)
GPT-4o (Dossier Extraction)	Metadata extraction	~$0.01-0.02	~$1-2
GPT-4o-mini/GPT-4o (Synopsis)	Synopsis generation	~$0.03-0.05	~$3-5
Claude Sonnet 4.5/GPT-4.1 (Script)	Script generation	~$0.15-0.25	~$15-25
Embeddings (text-embedding-3-small)	Story indexing	~$0.0001	~$0.01
Total per Story		**~$0.19-0.32**	**~$19-32**

Cost Breakdown Details:

- **Dossier Extraction:** ~500-1000 input tokens, ~200-500 output tokens = ~$0.01-0.02
- **Synopsis:** ~1000 input tokens, ~1500 output tokens = ~$0.03-0.05 (GPT-4o-mini) or ~$0.05-0.08 (GPT-4o)
- **Script:** ~2000 input tokens, ~6000 output tokens = ~$0.15-0.25 (Claude Sonnet 4.5) or ~$0.20-0.30 (GPT-4o)
- **Embeddings:** ~500 tokens per story = ~$0.0001

Infrastructure Costs

Service	Monthly Cost
Supabase (Free tier)	$0
Vercel (Free tier)	$0
Backend Hosting (Free tier)	$0
Total Infrastructure	**$0**

Total Monthly Cost

- **LLM API Usage:** ~$19-32/month (100 stories)
- **Infrastructure:** $0/month (all free tiers)
- **Total: ~$19-32/month**

Technical Implementation Details

Phase 1 Completion

Files to Modify:

- `stories-we-tell-backend/app/ai/dossier_extractor.py` - Add new metadata fields
- `stories-we-tell-backend/app/services/story_indexer.py` - New service for embeddings
- `Stories-we-tell/src/components/SidebarDossier.tsx` - Display new metadata

Database Migrations:

- `20250118000001_add_enhanced_metadata.sql`
- `20250118000002_add_story_embeddings.sql`

Phase 3: Synopsis Generation

Files to Modify:

- `stories-we-tell-backend/app/ai/models.py` - Add `_generate_synopsis_response` method
- `stories-we-tell-backend/app/api/simple_chat.py` - Generate synopsis in completion flow
- `stories-we-tell-backend/app/services/validation_service.py` - Store synopsis

Database Migration:

- `20250118000003_add_synopsis_to_validation_queue.sql`

Success Criteria

Phase 1 Completion

- ✅ All metadata fields captured in dossier (era, audience, emotional perspective)
- ✅ Story records indexed and searchable
- ✅ RAG layer functional for similarity search

Phase 3 Completion

- ✅ Synopsis generated from dossier when story completes
- ✅ Synopsis stored in validation queue
- ✅ Admin can view synopsis in validation detail page

Questions for Client

1. **LLM Preference:** Continue using Claude Sonnet 4.5 for scripts, or switch to GPT-4o?
2. **Synopsis Format:** 1-2 page narrative, or shorter summary format?
3. **Metadata Priority:** Which metadata fields are most important? (era, audience, emotional perspective)