

AI document processing

LLM cost estimation and batch-job checklist

From the field note "[LLM cost estimation for AI batch jobs: quotes that hold and runs that resume](#)".

The short version

- ◇ Reliable **LLM cost estimation** prices the run you will make: the real chunk plan, instructions once per chunk, and each pass separately.
- ◇ A quote holds when its inputs are hashed, it expires after 30 minutes, and its prompt and model versions are pinned.
- ◇ Chunks commit only after their output validates, so a failed job resumes where it broke and never repays finished work.
- ◇ Retries back off with jitter, honor Retry-After, skip errors a retry cannot fix, and reach a fallback model last.
- ◇ Long jobs survive crashes and deploys: stale workers cannot finish a job, orphaned jobs return to the queue, and deploys pause work.

The checklist

- ☐ Price from the real chunk plan, and size requests from the exact text sent.
- ☐ Keep 10% context headroom and reject oversized requests before any call.
- ☐ Hash every quote input, pin versions, and expire quotes after about 30 minutes.
- ☐ Commit each chunk only after validation, in one transaction with its status.
- ☐ Back off with jitter, honor Retry-After, and never retry authentication errors.
- ☐ Log tokens and cost per attempt, without storing customer text.
- ☐ Fence attempts, send heartbeats, and pause jobs before deploys.
- ☐ Set the broker visibility timeout above your longest job, and pair it with a reconciler.

Want a second pair of eyes on your answers?

Book a free 30-minute discovery call at www.seriustech.com/contact or email contact@seriustech.com. We reply within 1 business day.