Great question—those are **command-line flags** parsed by argparse so you can tweak behavior without editing code. Here’s what each does in:

python rag\_report.py -g --top-k 5 --max-words 250 --temperature 0.55 --model gpt-4

* -g / --use-general-knowledge  
  Enables **hybrid mode**: the model may use its own background knowledge **in addition to** your retrieved excerpts. (If you omit -g, it’s strict RAG: “excerpts only”.)
* --cite

Requires inline citation

* --top-k 5  
  Retrieve the **top 5** most relevant chunks from the index (default was 3).  
  Trade-off: more context = potentially better answers, but larger prompts (slower/costlier).
* --max-words 250  
  Chunk size during ingestion (how many words per chunk).  
  Smaller chunks = finer-grained retrieval but more chunks overall.  
  **Changing this triggers an index rebuild** (we check it via the manifest / corpus comparison logic).
* --temperature 0.55  
  Controls randomness/creativity of the model’s generation.  
  Lower (e.g., 0.2–0.4) = precise/consistent; higher (0.6–0.8) = more exploratory.  
  In our script, default is **0.4** for strict and **0.6** for hybrid unless you override it here.
* --model gpt-4  
  Which OpenAI chat model to use. You can swap to another compatible chat model name you have access to.

**Helpful notes about command-line flags**

* **Order doesn’t matter**: flags can appear in any order.
* **Defaults**: if you omit a flag, we use the script’s default (see argparse definitions).
* **Short vs long flags**: e.g., -g equals --use-general-knowledge.
* **Help anytime**: run python rag\_report.py -h to see all flags and defaults.
* **Persistence interaction**: changing --max-words, the **embedding model**, or the **documents** in input/ will cause an **auto-rebuild** of the FAISS index; otherwise we reuse it.

**When to change what**

* Increase --top-k if answers feel under-supported by context.
* Decrease --max-words if you want more pinpoint retrieval; increase it if chunks feel too fragmented.
* Use -g when you want comparisons to “generally accepted views” or broader synthesis.
* Tweak --temperature upward for broader, more discursive prose; downward for tight, factual output.

| **Short Flag** | **Long Flag** | **Type / Value** | **Default** | **Purpose** |
| --- | --- | --- | --- | --- |
| -g | --use-general-knowledge | Boolean switch | False | If set, runs in Hybrid Mode (LLM may use general knowledge + excerpts). If omitted, runs in Strict RAG mode (excerpts only). |
| (none) | --cite | Boolean switch | False | If set, requires inline citations like [Source p.X] after each factual claim, using only retrieved excerpts. |
| (none) | --top-k | Integer | 3 | Number of most relevant excerpts **kept** after FAISS retrieval (and optional re-ranking). |
| (none) | --max-words | Integer | 300 | Max words per chunk during ingestion. Changing this triggers index rebuild. |
| (none) | --model | String | "gpt-4" | Which OpenAI chat model to use for answering. |
| (none) | --temperature | Float (0–1) | 0.4 (strict), 0.6 (hybrid), 0.35 (strict+cite), 0.5 (hybrid+cite) | Controls creativity/randomness in output. Higher = more creative, lower = more factual/stable. |
| (none) | --expand-queries | Integer | 0 | Number of alternative phrasings of the user’s question to generate for expanded retrieval. |
| (none) | --per-query-k | Integer | 8 | Number of candidate chunks to retrieve per query/variant before merging and re-ranking. |
| (none) | --no-llm-rerank | Boolean switch | False (rerank ON by default) | If set, skips LLM re-ranking and uses FAISS similarity scores directly. |
| (none) | --rerank-model | String | "gpt-4o-mini" | Model used for the LLM re-ranking step. |