Step name	What it does
main_route	Top-level router: runs is_graph_query(query) and sends to either graph_chain or vector_chain.
main_route_kg_vs_vector	The routing decision: is_graph_query() (token/heuristic check for graph vs vector).
kg_chain	Full knowledge graph (kg) RAG pipeline: build context → augment sales → answer.
kg_build_context	Tool: Picks context type (person/org/sales) and builds knowledge graph context string; calls _graph_build_context().
kg_build_context (nested)	@traceable on _graph_build_context: same logic, appears as child when the tool runs.
format_org_summary	Builds org-chart summary string from the graph.
format_sales_overview	Builds Q3/sales overview string from the graph.
format_person_sales_summary	Builds sales summary for a specific person.
get_total_sales	Computes total sales (used inside format_sales_overview).
kg_augment_sales	Tool: For sales context, optionally adds product/marketing vector retrieval; calls _graph_add_vector_context().
kg_augment_sales (nested)	@traceable on _graph_add_vector_context: same logic as child.
kg_answer	Chain: Combines knowledge graph (and optional vector) context, runs prompt + LLM + parser, returns {answer, sources}.
kg_answer (nested)	@traceable on _graph_answer_llm: same logic as child.
kg_generate	Prompt → LLM → StrOutputParser for the final graph answer.
ChatGoogleGenerativeAI (or model nam	LLM call for the knowledge graph answer.
vector_chain	Full vector RAG pipeline: classify → retrieve → format → answer.
vector_classify	Chain: Keeps query and runs document-type classification in parallel.
vector_classify_document_type	Prompt → LLM → JsonOutputParser to pick category (HR Policy, Marketing, etc.).
ChatGoogleGenerativeAI (or model nar	LLM call for classification or for vector answer.
vector_retrieve_step	Chain: Builds filter from classification, invokes retriever with config, returns {query, docs}.
vector_retrieve	Retriever: Fetches documents from the vector store (optional filter by category).
vector_format_context	Chain: Turns {query, docs} into {query, context, sources} for the LLM.
vector_answer	Chain: Runs answer generation and passes through sources in parallel.
vector_generate	Prompt → LLM → StrOutputParser for the vector RAG answer.
ChatGoogleGenerativeAI (or model nar	LLM call for classification or for vector answer.
chitown_chat_orchestrator	Top-level @traceable wrapper around chat_chain.invoke(user_query).