How could you use AI to create an optimized context so that user does not need to repeat what he said previously into conversation with AI?

You can follow one or any of these alternatives:

- Create own architecture which manages context. This architecture may contain algorithms that decide what user message (old one) is better related to AI response and vice-versa, or indeed, what set of user messages (old ones) are better related to AI response and vice-versa. This can provide a much smart and quick interaction during conversation between user and AI.

- Generate extra data by analyzing semantically each user message and eventually, you can use this additional data to know better how to respond to the current user message.

- Discard irrelevant user messages to avoid response creation by AI. This may reduce significantly the delivery time response.