Une fois toutes ces parties gérées, il faut bien répondre à l’utilisateur, c’est le sens même d’un Chatbot. Pour ce qui concerne notre Chatbot, nous l’avons évoquée tout a l’heure, ce sera des réponses fixes et aléatoire. Pour chaque *intent*, nous aurons un certain nombre de réponses qui seront associée, c’est pour rendre plus agréable l’expérience de l’utilisateur. Comment marche les réponses fixes mais aléatoire ? Dans le cas d’une salutation par exemple, nous allons avoir 3 réponses comme suit : « bonjour comment allez-vous », « bien le bonjour mon très cher », « je vous salue ». A partir de là, une fois une salutation détectée, nous lui renvoyons une de ces trois (3) réponses au hasard.

Maintenant, ce subterfuge ne suffit pas vraiment pour convaincre de « humanité » du Chatbot a l’utilisateur. Car a un moment il va bien se rendre compte que ce dernier ne renvoie que les mêmes réponses, même cela ne constitue pas un problème pour ce genre de Chatbot. C’est le lien d’introduire une des éléments les plus important dans le développement de Chatbot actuellement, ce qui se fait de mieux à l’heure où j’écris ces lignes : ce sont les *Large Language Model* (LLM). Dans ce document nous avons vu comment faire un *Intent classification* qui est moyen de faire passer du texte dans un model de ML, c’est ce qu’on appelle aussi dans le jargon un *Language Model* (LM), c’est-à-dire de la classification de texte. Maintenant, et c’est que cela devient intéressant, notre dataset se constituait d’à peu près 1000 lignes étiquetés, d’accord. Imagions juste que nous avons un dataset qui se compose de l’ensemble des informations sur internet (ce qui est juste énorme) ou un qui se rapproche d’une telle abondance de données, là on parle de LLM. Il y a beaucoup de LLM aujourd’hui dont GPT, Gemine, Llama etc…

Il était opportun de faire ce détour car nous allons utiliser un LLM en l’occurrence Llama (LLM gratuit de Meta) qui va nous aider dans la gestion des réponses. Comment ? LLama va nous permettre tout simplement de formuler les réponses pour qu’elles soient moins redondantes. Ce LLM pourrait aussi nous permettre d’optimiser le *Intent classification.* Le terme utilisé pour communiquer avec un LLM a travers du code est …