



# GRILLBot In Practice: Lessons and Tradeoffs Deploying Large Language Models for Adaptable Conversational Task Assistants

Sophie Fischer\*  
University of Edinburgh  
Edinburgh, United Kingdom  
s.fischer@ed.ac.uk

Carlos Gemmell  
University of Glasgow  
Glasgow, United Kingdom  
c.gemmell.1@research.gla.ac.uk

Niklas Tecklenburg\*  
University of Glasgow  
Glasgow, United Kingdom  
tecklenburg.niklas@gmail.com

Iain Mackie  
University of Glasgow  
Glasgow, United Kingdom  
i.mackie.1@research.gla.ac.uk

Federico Rossetto  
University of Glasgow  
Glasgow, United Kingdom  
2507743r@student.gla.ac.uk

Jeffrey Dalton\*  
University of Edinburgh  
Edinburgh, United Kingdom  
jeff.dalton@ed.ac.uk

## ABSTRACT

We tackle the challenge of building real-world multimodal assistants for complex real-world tasks. We describe the practicalities and challenges of developing and deploying GRILLBot, a leading (first and second prize winning in 2022 and 2023) system deployed in the Alexa Prize TaskBot Challenge. Building on our Open Assistant Toolkit (OAT) framework, we propose a hybrid architecture that leverages Large Language Models (LLMs) and specialised models tuned for specific subtasks requiring very low latency. OAT allows us to define when, how and which LLMs should be used in a structured and deployable manner. For knowledge-grounded question answering and live task adaptations, we show that LLM reasoning abilities over task context and world knowledge outweigh latency concerns. For dialogue state management, we implement a code generation approach and show that specialised smaller models have 84% effectiveness with 100x lower latency. Overall, we provide insights and discuss tradeoffs for deploying both traditional models and LLMs to users in complex real-world multimodal environments in the Alexa TaskBot challenge. These experiences will continue to evolve as LLMs become more capable and efficient – fundamentally reshaping OAT and future assistant architectures.

## CCS CONCEPTS

• **Computing methodologies** → **Natural language generation; Discourse, dialogue and pragmatics.**

## KEYWORDS

Conversational Task Assistants, Large Language Models

### ACM Reference Format:

Sophie Fischer, Carlos Gemmell, Niklas Tecklenburg, Iain Mackie, Federico Rossetto, and Jeffrey Dalton. 2024. GRILLBot In Practice: Lessons and Tradeoffs Deploying Large Language Models for Adaptable Conversational Task Assistants. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge*

\*Work done at University of Glasgow.



This work is licensed under a Creative Commons Attribution International 4.0 License.

KDD '24, August 25–29, 2024, Barcelona, Spain  
© 2024 Copyright held by the owner/author(s).  
ACM ISBN 979-8-4007-0490-1/24/08  
<https://doi.org/10.1145/3637528.3671622>

Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.  
ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3637528.3671622>

## 1 INTRODUCTION

Interactive assistants based on Large Language Models (LLMs) [20, 30, 32] are rapidly evolving and changing expectations for their fluency and new capabilities. LLMs trained for chat and dialogue produce fluent textual responses and can track state for multiple turns. However, practically deploying them at scale with voice-based assistants like Alexa, Siri, and Google Assistant remains challenging. We further envision agents that assist people in complex long-running real-world tasks requiring a high degree of reliability, such as cooking dinner menus, refurbishing the kitchen, or taking a personalised at-home workout routine. These tasks represent fundamental challenges because they are grounded in real-world tasks and need to adapt to a changing and dynamic environment. Existing end-to-end dialogue systems solely based on LLMs are currently not capable of this since they can hallucinate and struggle to maintain context [14, 34].

In this work, we address this gap by presenting new generative AI methods that underlie the online GRILLBot Alexa Prize system that won the first and second prizes in the Alexa TaskBot Challenge [9, 10]. GRILLBot assists people with real-world problems at home, such as cooking and other physical tasks, and is battle-tested by hundreds of thousands of users across the US over multiple years and generations. Fig. 1 shows a simplified example cooking conversation with GRILLBot.

From the beginning, GRILLBot built on generative language models to be flexible and adaptable. Instead of end-to-end generation, it leverages a hybrid approach that uses specialised models to handle specific tasks. The decision to deploy LLMs for (some) of these models has important tradeoffs that need to be managed carefully. We present lessons and challenges deploying GRILLBot with hard constraints on response latency, reliability (uptime), and compute resources with the need to continuously handle concurrent conversations from Alexa users.

We leverage LLM utility to provide a rich and engaging user experience with unique and differentiating capabilities for the Taskbot. For example, GRILLBot preprocesses task data offline [9] and online to respond to a dynamic user environment. We detail the challenges and tradeoffs of deploying LLMs versus specialised models in the



**Figure 1: A multimodal conversation with OAT including task adaptation and question answering with system actions by the NDP in green.**

key online components that make conversations effective. Key components covered include 1) generating flexible system actions from code generation, 2) responding to dynamic information needs with knowledge-grounded question answering, and 3) modifying tasks online to adapt a task to the user's preferences and constraints.

First, we discuss and evaluate GRILLBot's unique approach to handling diverse and dynamic interaction patterns in complex task-oriented conversations. Instead of traditional intent-classification models [2, 12, 16], GRILLBot defines a Neural Decision Parser (NDP) model that acts as a system orchestrator. Given the dialogue history, it generates actions the system should take in the form of generated code in an extensible domain-specific language. We show that a specialised supervised NDP model learned from a small set of a few hundred carefully curated examples can outperform much larger models requiring significantly more data. Experiments show that for this critical and latency-sensitive component triggered on all interactions, a specialised model provides a 100x latency advantage.

A robust knowledge-grounded Question Answering module is the second key element that allows the system to react robustly to unpredictable users. A TaskBot QA system needs to reason across task and conversation history to handle dynamic information requests. LLM-based systems are more capable of this than traditional extractive QA models, but standard generative QA is not grounded in a task and ongoing environment. In addition, current existing conversational datasets [3, 21, 25] fail to simulate real-world task-oriented conversations. To fill this gap, we extend the Wizard-of-Tasks (WoT) task-oriented conversation benchmark [5] to a new

task-oriented QA dataset using further web crawling and manual annotation. We experiment with QA models and LLMs to perform contextualised task-grounded question answering. Human annotators agree that LLMs respond more correctly to abstractive QA. However, advanced neural models like Unified QA [15] outperform LLMs for extractive QA according to human annotation, F1 score and latency.

Finally, we leverage the world knowledge from LLM pretraining to modify tasks according to a user's preferences and constraints. To focus on correctness and safety, we use a hybrid approach for live task adaptation. If a user requests a change to the task, e.g. changing a pizza to be vegetarian, we call an LLM-based task rewriter that takes the current task and adapts it to the user's constraints. The rewriter outputs the task in a structured format (JSON) so that the system framework can access the edits for the remaining conversation. With manual annotation, we show that our task rewriter managed to adapt a task successfully in 56% of cases. Of the successful adaptations, 73% of suggested LLM replacements were sensible and would work in the real world. This shows the potential of using LLMs for structured editing of underlying data structures in a hybrid assistant architecture.

Our contributions are:

- We describe the GRILLBot online architecture and lessons and insights on developing its hybrid design leveraging both LLMs and specialised models for key components.
- For contextualised task-focused QA, we extend the WoT benchmark dataset [5] to evaluate correctness and groundedness for complex task questions, creating a new dataset *WoTe* that we release publicly. We evaluate neural and LLM models available during the TaskBot challenge, showing that LLMs beat neural models in abstractive QA, but are outperformed for extractive QA.
- For system orchestration and dialogue management, we perform experiments with the Neural Decision Parser (NDP). Results show smaller, specialised language models are highly adaptable and have high effectiveness with 100x lower latency.
- We study the effectiveness of LLM-based edits to tasks. Results show that the LLM's real-world knowledge and fluency enable structured changes to underlying data structures, with 73% of replacements being sensible.

Throughout this paper, we share key insights into user behaviour and lessons learned deploying GRILLBot to thousands of users with models refined and developed across multiple years of the Alexa Prize TaskBot Challenge. GRILLBot was one of the first to adopt LLMs online for complex task responses. GRILLBot is reproducible with all non-user data and key components released continuously in the OAT framework [11], which we base GRILLBot on. The continued evolution of best practices during deployment holds important lessons for both the current and future task assistants and their use of generative LLMs.

## 2 RELATED WORK

### 2.1 End-to-end dialogue models

End-to-end dialogue models based on transformers are fine-tuned on chat data and use LLM generation without underlying specialised modules. Models like LaMDA [32], BlenderBot 3 [30], and WikiChat [29] benefit from model scaling to generate higher quality responses. However, many leverage proprietary data and are not publicly available for custom assistants (e.g. [20, 32]).

In contrast to chat models, TaskBots require task-oriented conversations that are longer and more specialised. When a TaskBot guides the user through the task, the assistant leads the conversation. TaskBots are more proactive and react flexibly to requests to actively shape the underlying task. However, most task-oriented datasets are user-led and the user asks the assistant to perform a task, like booking a hotel. Conversations are oriented on pre-defined slot-filling conversation flows [3]. In comparison, the Wizard of Tasks (WoT) dataset [5] contains conversations between crowd workers acting as students and teachers within the cooking and DIY domains. This means we have nearly no training data for training models for the TaskBot task and start in a low-resource setting during development.

### 2.2 Modular Agent Architectures

Compared to end-to-end models, modular conversational agents split control over system behaviour into specialised components like response generation, retrieval and dialogue management. Different conversational agent frameworks have been created to help with boilerplate code to provide building blocks of agents.

Popular frameworks [2, 12] are not fine-grained and flexible enough to allow specialised model and LLM deployment. To enable fine-grained control of model use and own hosting rights, various research institutions publish their frameworks [4, 11, 22, 38], some of which built and battle-tested during Alexa Prize Challenges. We base GRILLBot on the public open-source OAT framework [11] developed over multiple years of the TaskBot challenge. OAT is a modularised task-oriented conversational agent framework which achieves a scalable, lightweight, and non-resource-intensive architecture with low latency.

**2.2.1 State management.** Dialogue State Tracking (DST) is a standard task in task-oriented conversational agents. Traditional approaches include a model conversation with predefined schemas that structure dialogue into intents and slots [36]. Google [12], Amazon [2] and RASA [26] follow this approach. Usually, dialogue management leverages hierarchical state machines or flow controllers [1]. Since intent flows are fixed and once flow is selected, navigating away is complicated, intent models are very brittle.

Gemmell et al. [10] instead introduce NDP models that generate flexible system actions. We build upon this work by evaluating different models to perform the NDP task. We explore how bigger models with zero-shot/ few-shot or in-context learning perform and discuss tradeoffs in latency and effectiveness.

### 2.3 Task-specific question answering

Previous work shows that generative models performing long-form QA tend to add additional information or hallucinate answers [14].

This is potentially dangerous in a real-world setting and can undermine the agent’s perceived trustworthiness in the limited user interaction [28].

Khashabi et al. [15] convert the task context into a natural language representation and pass it into a model jointly with the user query. Lewis et al. [17], Shuster et al. [30] combine this approach with retrieving from relevant dialogue context. Choi et al. [5] experiment with abstractive question answering with fine-tuned versions of BART and T5. Their models hallucinate numerical terms and units and show low performance, showcasing the challenging task. Since models like T5 can only ingest a limited context length, we implement pre-processing of context to shorten to the available token length. In comparison, in-context learning with few-shot prompts of LLMs needs limited training data. Context length is less restricted, but the length of the generative decoding and model size are computationally more expensive and add extra latency.

In this work, we trial both traditional and LLM models as a basis for abstractive and extractive QA tasks. To ensure knowledge grounding of target answers, inspired by Khashabi et al. [15], Rajpurkar et al. [25], we reformulate the QA task to be extractive. Given a context paragraph and a question, the model needs to extract the answer from the paragraph by selecting a substring. We follow Zaib et al. [35] by classifying questions into factoid, causal, confirmation, listing, and complex questions to allow further fine-grained analysis.

### 2.4 Dynamic Task Adaptation

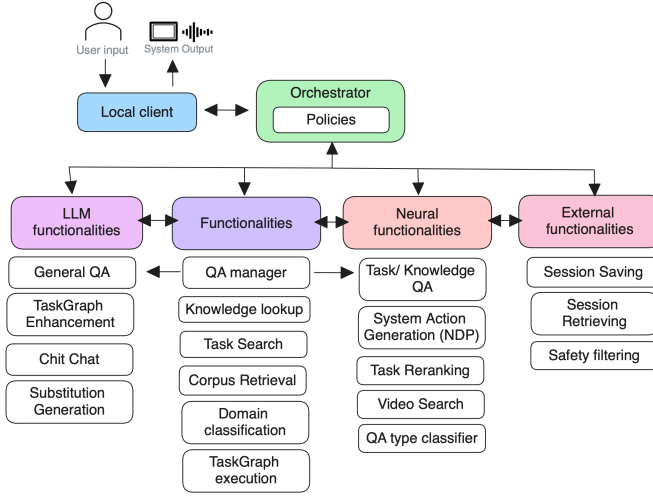
Due to the dynamic nature of real-world tasks, a virtual assistant needs to be able to listen, understand, and adapt the task based on the user’s input. OAT represents tasks as TaskGraphs, which allows dynamic editing and scheduling of task components [11]. In previous work, we perform task augmentations offline to create more engaging conversations, including non-linear conversations [10], adding additional details, splitting steps, and writing task descriptions [9], and aligning videos [8].

However, it is impossible to predict all possible live user requests before task execution. In this work, we therefore use LLMs to edit the TaskGraphs live during the conversation, so that the system responds to unforeseen information and modifies tasks. One example of this is substituting ingredients for recipes and adapting the task based on user preferences. Various approaches span using templates and external knowledge sources [7] to training specialised models [18, 19, 23].

## 3 IMPLEMENTATION DETAILS

### 3.1 TaskBot Task

In this section, we define the TaskBot task more formally. Given a conversational history  $[C_1, \dots, C_n]$ , we find an explicit matching Task  $T$  that the user would like assistance such as cooking a recipe or refurbishing the kitchen. Then, we guide the user through  $T$  by scheduling step-by-step actions  $[S_1, \dots, S_n]$  dynamically. When managing the dialogue and responding to users, at each response we consider the task  $T$  and conversational history  $[C_1, \dots, C_n]$  when generating the system response  $R$ . There are no explicit conversational flows, meaning that the system can flexibly react to user requests at any time of the conversation.



**Figure 2: Online architecture of GRILLBot based on OAT [11]. We implement NDP (Section 3.3) & QA (Section 3.4) in Neural functionalities and task adaptation in (Section 3.5) in LLM functionalities.**

### 3.2 Online GRILLBot System Architecture

Fig. 2 shows the different modular components of our deployed online TaskBot system built on the OAT framework. Using the Orchestrator module, we create several policies for GRILLBot which handle different functionalities grouped by resource requirements. *LLM functionalities* contain all generative capabilities. We create features for general QA, chit-chat, and various conversation enhancements such as TaskGraph adaptations. *Neural functionalities* handle all neural models requiring GPU, such as system action generation and task reranking. *Main functionalities* include features for retrieval, lookup and domain classifications. During the development of GRILLBot, we continue releasing models and synthetic training data as part of the OAT framework [11].

GRILLBot uses a Docker and Kubernetes setup to manage resources and maintain constant response times. Docker allows running the entire application by only installing the single Docker dependency and no virtual environments. Modular Docker containers help with version control, installing dependencies, and decoupling. Resource-heavy components that host neural models and LLMs do not interrupt more lightweight components. This setup helps adapt to traffic and usage spikes and maintain low latency. Since GRILLBot is a live system, we have explicit latency constraints for modular components. We aim to give answers in less than 1.5 seconds, which we manage in over 93% of utterances. Battle-testing GRILLBot with thousands of users, Kubernetes successfully managed load-balancing system components with an average system latency under 0.5 and 1.1 seconds.

### 3.3 Code generation for dialogue management

To overcome the brittleness of traditional intent classification, in previous work, Gemmell et al. [10] create Neural Decision Parsers (NDP) to generate code to represent system actions. We define the

code generation task for managing dialogue as follows: Given a Task  $T$  and conversational history  $\{C_1, \dots, C_n\}$  represented in natural language as input sequence  $\{x_1, \dots, x_n\}$ , auto regressive generate system action  $a = \{y_1, \dots, y_m\}$ .

In theory, the action space that includes all  $a$  is unlimited. However, since we can only execute supported actions by the system back end, we fix the action space  $A = \{a_1, \dots, a_j\}$  for practical reasons to represent available system capabilities depending on the training data. The NDP can generate action arguments freely, such as the search arguments in *search("veggie pizza")* or the selection option in *select(1)*. All actions not in  $A$ , i.e. beyond system capabilities, are handled by a Fallback LLM to generate a fluent response without executing any system actions. We ensure that the LLM Fallback does not hallucinate by adding clear constraints in the LLM prompt of what system capabilities are. We also ask the model to ask polite questions, if the user request is unclear. In addition, we leverage the Alexa Prize CoBot system's [16] safety classifiers to ensure no dangerous responses are generated.

Fig. 1 shows examples of NDP output in green. The NDP translates conversational state, history, and task state into appropriate system actions. This enables the system to parse the user request flexibly. The deployed NDP model has a strict time constraint of  $< 0.2$  seconds since the system calls the NDP model at each conversation turn and follow-up calls need budget to execute under constraints.

In Section 5, we experiment with different model sizes and types, such as encoder-decoder models versus decoder-only models. We also compare bigger models versus finetuning on a custom-curated benchmark and discuss tradeoffs in effectiveness and latency. For deployment, we use small specialised supervised NDP models trained from small seq-2-seq models [6, 24] learned from a few hundred carefully curated examples, which have high accuracy and low latency.

### 3.4 Task-specific retrieval-augmented question answering

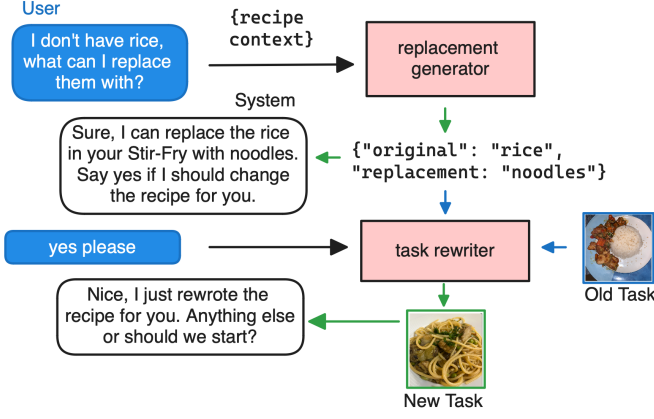
Using the NDP, the system can identify when the user asks a question and forward this to the system's specialised question-answering module. We define the task-oriented question-answering as follows: Given a Task  $T$  and conversational history  $\{C_1, \dots, C_n\}$  (with the user question being  $Q = C_n$ ), generate a system response  $R$  that answers  $Q$ .

In the QA module, we pass the user question, the most relevant task context, and conversation history into the model. For different QA types, GRILLBot uses neural extractive [6, 15] and LLM-based QA approaches [31, 33] to generate relevant answers based on the passed context. Neural and LLM-base approaches have different advantages in latency, computational resources needed and model abilities. In Section 5, we discuss tradeoffs between different model types for abstractive and extractive questions.

### 3.5 Live generative task adaption

A flexible task assistant needs to be able to adapt a task based on user utterances and preferences. We define Task Adaptation as follows in two steps. First, given a user replacement question  $Q$  and the current Task  $T$ , we identify the original requirements  $[O_1, \dots, O_n]$





**Figure 3: Live task adaptation based on the Replacement Generator and Task Rewriter.**

to replace and map them to new requirements in a replacement mapping  $O_1 : R_1, \dots, O_n : R_n$ . Requirements can be ingredients or tools the user needs for  $T$ . Second, given the mapping of old to new requirements, we rewrite the Task  $T$  with instructions  $[S_1, \dots, S_n]$  and original requirements  $[O_1, \dots, O_n]$  to create  $T'$  with rewritten steps  $[S'_1, \dots, S'_n]$  and rewritten requirements  $[R_1, O_2, \dots, R_n]$ .

To perform the task adaptation, we build an LLM-based substitution engine that allows modifying task ingredients, tools, and task steps live to suit the user's replacement request. Fig. 3 shows a conversation with background LLM calls. If the NDP detects a user substitution request, the system queries the LLM with a replacement request in a pre-defined prompt with a filled-in context. If the user replacement request is valid, the LLM offers to replace the old with the new replacement and rewrite the task to reflect the changes. If the user agrees, we select the steps and ingredients that need replacing. For each step and ingredient in the replacement mapping, we prompt the LLM to perform an edit.

#### 4 LESSONS LEARNED AND SHORTCOMINGS

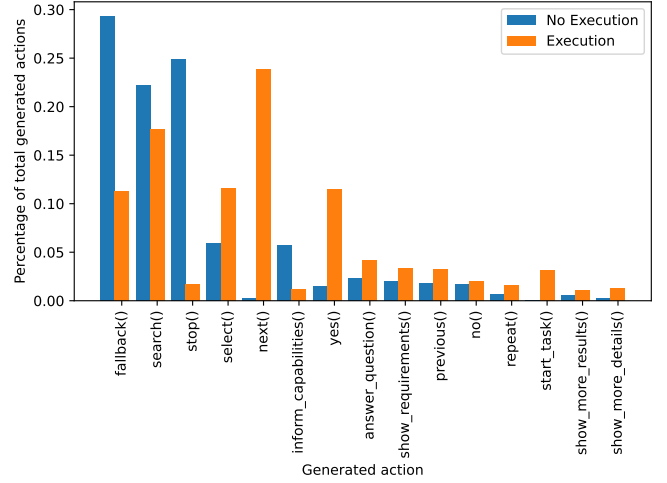
When we started developing GRILLBot in 2021, few openly accessible live virtual assistants using generative models existed. We build our OAT framework to allow scalable modular components to support using models live, which works well with Docker and Kubernetes deployment. GRILLBot keeps the daily average latency under 1.1 seconds despite the high traffic of thousands of users.

During the journey of developing GRILLBot, we explore tradeoffs for using LLMs within a live system. An example of this is the NDP. When we started the challenge, the NDP was a basic T5 model trained on a few hundred hand-crafted training examples. For us, it is remarkable how small sequence-to-sequence models still manage to keep up with few-shot in-context learning of models with many more parameters. This allows us to keep the latency of the frequently called NDP low and shows that LLMs might not always be the answer.

This becomes especially important when we start chaining models. Balancing the cost of resources and improvements in performance is increasingly difficult. Table 1 shows the latency of a few

**Table 1: Latency of selected system features before Week 29 and after, when we deployed LLMs in the system.**

Action	Latency (in sec)		% increase
	before LLMs	with LLMs	
fallback()	0.54s	1.14s	114%
answer_question()	0.89s	1.66s	87%
search()	0.92s	0.94s	2%
replace()	-	2.38s	-



**Figure 4: Generated action distribution from conversations where a task is started compared to exploratory-only.**

selected system components. In Calendar Week 29, we start deploying larger LLMs in the live system, which increases latency. For various system features, we start calling the LLM endpoint deployed on a single NVIDIA A10G GPU with 24 GiB memory. We zero-shot prompt the model with action-specific handcrafted prompts and contexts. Fallback and QA use one generative call, whereas the task adaptation engine chains two generative calls.

As a result, fallback response and answer question times double compared to our previous approach of using lightweight finetuned encoder-decoder models for fallback and QA. Since we set the maximum time for LLM generation to 2 seconds, 1.7% of fallback and 19% of question answer actions time out and the system responds with a few standard default responses. However, since the deployment of the LLM, user ratings of conversations increase by 13%. Conversations with questions and fallback see an increase in user ratings of 30% and 10%, respectively.

During log analysis, we review conversations after Week 29. Fig. 4 shows types of action codes generated by the live NDP over the entire span of the competition, highlighting how many utterances are handled to the LLM fallback. Most utterances are chit-chat requests, highlighting that users love chatting with assistants and trying to break them. We observe that more than 30% of user utterances are handled by the LLM fallback if the user is not in execution,

**Table 2: Evaluation of NDP models. We finetune all models on the train split.**

Model	Accuracy	Precision	Recall	F1	Latency
t5-base [24]	0.823	0.602	0.571	0.575	0.01s
unifiedqa-t5-base [15]	0.785	0.558	0.506	0.515	0.01s
flan-t5-base [6]	0.839	0.608	0.566	0.574	0.01s
llama-2-7b [33]	0.881	0.739	0.688	0.710	1.22s
llama-2-7b-chat [33]	0.203	0.114	0.132	0.420	2.12s

i.e. if the user has not chosen a task. Therefore, handling those user requests flexibly and fluently is most important.

One of our biggest lessons learnt is that LLMs are not the answer for every single system component. For components with low latency requirements, finetuning specialised models is more sensible. For system components requiring fluent and complex responses such as QA and fallback, deploying a LLM in a structured manner is effective for answer quality despite increased latency since users are very unpredictable. In addition, we learn that we don't need to finetune expensive models for fluent response generation if we carefully prompt the model with the right context and implement safeguards.

## 5 EVALUATION

To decide which models to use for system components, we perform component-level evaluation. We compare which models can be trained on system action code generation to accurately translate user utterances into executable system actions. Then, we evaluate which models perform best at both abstractive and extractive task-specific question answering. Finally, we review the performance of the task adaptation feature.

### 5.1 NDP evaluation

**5.1.1 Dataset Creation.** Building on previous work [9], we extend the existing dataset by 25% with rewritten user logs and additional synthesised logs to test action code generation with different NDP model versions.

This test set includes user utterances with previous system responses, predicted intent by the system, and a correct intent prediction annotation. We split this test set into 60% training, 10% validation, and 30% testing to ensure an even user request distribution during testing. We can't disclose the amount of original user log data, but we generate synthetic data by prompting ChatGPT [20] to balance the intent distribution. We release the synthetic NDP training data as part of the most recent OAT release[11]<sup>1</sup>.

**5.1.2 Metrics and Baselines.** We fine-tune various encoder-decoder models such as UnifiedQA [15], T5 [24], FLAN T5 [6] as well as the decoder-only Llama 2 base model (Llama-2-7b-hf) [33] on the test split. We train all models for one epoch on one machine with one NVIDIA A10G GPU with 24 GiB memory. We calculate precision, accuracy, recall, and F1 score averaged over all data. We also report average latency per action code generation.

**5.1.3 Results.** Table 2 shows model effectiveness on the test dataset. Llama 2 outperforms all models. We also finetuned the Llama 2 chat version, but it does not follow the action code input format well (36% of generations are non-parseable), as reflected by the evaluation metrics. For further insights, we compare our best-performing encoder-decoder (FLAN-T5) model to our best-performing decoder-only (Llama 2) model. FLAN-T5 follows the input format better than the Llama 2 during generation. 1.2% of generations with the FLAN-T5 do not match the possible action target space, compared to 11.8% with Llama 2.

Reviewing individual answers, Llama 2 is better at complex reasoning compared to FLAN-T5, which produces more wrong action codes. This is reflected by the F1 score of the Llama 2 model, beating the other baselines by a large margin of ~0.15. However, Llama 2 does not handle uncertainty well. Especially when the user is vague during navigation, option selection or task searching, the model hallucinates vague responses so that our LLM fallback handles the response. An example of this is co-reference. The user asks “go to the step after please” after having heard Step 1. The correct answer is “step\_select(2)”. Llama 2 instead generates “(step\_select, unknown)” which is a non-parseable wrongly formatted action code, which means that no system action is executed. Overall, T5 models have the advantage of a 100x lower latency compared to Llama models and we therefore prefer them for the repetitive calling of the NDP in the live system.

### 5.2 Task-specific QA evaluation

**5.2.1 Dataset Creation - WoTe.** As our QA test dataset, we extend the Wizard-of-Task dataset [5]. Due to its conversational task-oriented user-lead nature, this dataset is closest to a real conversation with a TaskBot. The original dataset contains ~17000 utterances from various conversations within the cooking and DIY domains. We filter out non-question user utterances by crowd worker annotation. This results in 4351 question-answer pairs, of which we keep 1589 which are answerable with the task context. Next, we drop all questions labelled irrelevant and not useful by crowd workers, resulting in 1337 questions. The original dataset does not include the task content, only links to task websites, which we need for factually grounding answers. We scrape task content of linked tasks, which is successful for 83% of tasks (1109). We also remove pairs with inconsistent labels which require common or external knowledge. This results in 827 final questions.

We then manually annotate the remaining questions by adding the extractive span that answers the question. We use the guideline of selecting the first occurrence of the answer within the context and keeping the answer span as short as possible.

We also add a taxonomy to classify questions more granularly following Zaib et al. [35]. We add two extra categories to the existing five (factoid, causal, confirmation, listing, complex). The *History* category describes questions where users ask for repetition from the conversational context. *Navigation* describes questions that ask the teacher to navigate through the task, i.e. moving forward a step. Table 3 shows the distribution of types within the dataset and example questions. We release the resulting *WoTe (Wizard of Tasks - extractive)* dataset on GitHub<sup>2</sup>.

<sup>1</sup><https://github.com/grill-lab/OAT>

<sup>2</sup><https://github.com/grill-lab/WoTe>

**Table 3: WoT(e) dataset question category distribution**

Question type	Count	Example Question
Factoid	276	Can the almonds be roasted or do they need to be raw?
Navigation	146	Once the fill tubing is installed, what step comes next?
Confirmation	131	Would my kitchen windowsill be a good place for the onions?
Complex	82	Does that mean basil grows best in the spring and summer?
Causal	50	Why shouldn't I mix in the sour cream at the same time?
History	33	Sorry, what do I need to do?
Listing	27	How much cream cheese and other ingredients will I need?

**5.2.2 Metrics and Baselines.** We compare traditional neural QA models such as FLAN T5 [6], UnifiedQA [15] and T5 [24] with generative LLM models such as Llama 2 [33]. We use off-the-shelf models that can run on a single GPU with minimal tuning for evaluation. To ensure even distribution on the rather small data set, we employ a 30% train, 20% validation, and 50% testing split. We finetune the models on the train split on an NVIDIA A10G GPU with the training objective of minimising the loss function of predicting the start and end token of the answer span.

For T5 models, we concatenate the tokens of question  $Q$  and context  $C$ . Due to limited token length, we retrieve the most relevant step using sBERT [27] for the T5 models. We also add a gold context baseline where we manually create the context to ensure the correct answer is included in the context. We follow related work [5, 25] and report SQuAD token-wise metrics and ROUGE and BERT-Score [37]. We evaluate the effect of fine-tuning and compare in-context learning to transfer learning.

**5.2.3 Abstractive Question Answering.** Choi et al. [5] provide original answers by human crowd workers as target answers in WoT. We experiment with more advanced generative models than the provided baselines by the authors for the abstractive QA task.

For T5 models, we notice issues with context parsing during implementation. We shorten the passed context to the most relevant step for most inputs to stay beneath the maximum input token length. However, on our test set, our automatic truncation using sBERT out-of-the-box only extracts the correct response of 45% of samples (Precision = 0.54, Recall = 0.23). sBERT fails when reasoning is required to select the step, for example, to answer a complex question that requires combining steps. Another failure point is questions that contain many words from another step, e.g. if a user rephrases a step as part of their question. Furthermore, navigational questions that require selecting a specific step are difficult. Therefore, in further evaluation, we only use manually annotated context to ensure the correct answer is in the context to ensure fair model comparison.

Table 4 show model effectiveness for abstractive QA. All models perform badly with Rouge scores < 0.3. We verify this by manually annotating 50 random questions to evaluate model performance for

**Table 4: Abstractive QA task evaluation. \* = finetuned, bold = significantly different to pre-trained T5 baseline**

Model	Rouge1	EM	F1	BERT-s
Llama-2-7b-hf	0.154	0.000	0.127	0.749
Llama-2-7b-chat-hf	0.230	0.000	0.198	0.863
t5-base *	0.267	0.000	0.237	0.874
unifiedqa-t5-base *	0.273	0.002	0.238	0.878
flan-t5-base *	<b>0.290</b>	0.000	0.252	<b>0.880</b>
Llama-2-7b-hf *	0.225	0.000	0.195	0.865
Llama-2-7b-chat-hf *	0.236	0.000	0.206	0.866

**Table 5: Abstractive QA manual annotation. Each answer is labelled between 0-2 by expert annotators.**

	Correct	Complete	Understandable
t5-base *	0.88	0.88	1.04
unifiedqa-t5-base *	<b>1.2</b>	1.1	<b>1.32</b>
flan-t5-base *	1.00	0.94	1.06
Llama-2-7b-hf *	<b>1.36</b>	<b>1.32</b>	<b>1.56</b>
Llama-2-7b-chat-hf *	<b>1.42</b>	<b>1.34</b>	<b>1.50</b>

correctness, completeness and understandability on a scale from 0-2 (0: not, 1: somewhat, 2: fully). Table 5 shows manual annotation results. We observe that annotators disagree with the metrics performance. Especially for generative models, annotators agree that almost always mostly or fully correct, completely understandable and significantly better than the t5 baseline.

We investigate why there is a discrepancy between user ratings and metrics. Comparing model and teacher answers, teacher answers in the original dataset are often noisy. Teachers omit task details required (e.g. *food-2-1*, *food-10-0*, *food-135-4*), could have answered from the task context (*food-51-1*), or are simply wrong (*food-44-8*, *diy-194-8*). Since this phenomenon repeats itself for many questions, the original answers are unusable for the task evaluation.

**5.2.4 Extractive Question Answering.** For more accurate evaluation, we change the QA task to be extractive. We use our annotated extracted answer snippets from the task context and conversation history and compare model output to the factually grounded context snippets. We define the extractive QA task as follows. Given a user question  $Q$  and a conversational context  $C$ , the model extracts the answer substring from  $C$ . The conversational context  $C$  contains information about the task, such as task title, description, steps, and ingredients/ requirements.

Table 6 shows different model performance on the task-oriented extractive QA task. We compare zero-shot and finetuned models with t5 base zero-shot and finetuned as baselines, respectively. Across the board, all models perform badly with low metric scores. In addition, compared to the finetuned T5 baseline, none of the models perform significantly better. The two generative models, Llama 2 and Llama 2, even chat perform worse than the baseline. To verify those results, we annotate 50 random questions and each model's outputs on a scale from 0-2 for correctness and completeness (0: not,

**Table 6: Extractive QA results. \* = finetuned, bold = means significant compared to baselines t5 base/ finetuned t5 base**

Model	Rouge	EM	F1	BERT-s
t5-base	0.126	0.022	0.117	0.437
unifiedqa-t5-base	<b>0.236</b>	<b>0.068</b>	<b>0.221</b>	<b>0.553</b>
flan-t5-base	<b>0.203</b>	0.034	<b>0.179</b>	<b>0.547</b>
Llama-2-7b-hf	0.146	0.015	0.127	0.458
Llama-2-7b-chat-hf	<b>0.237</b>	0.007	<b>0.224</b>	<b>0.575</b>
t5-base *	0.444	0.194	0.428	0.695
unifiedqa-t5 *	0.453	0.180	0.440	0.696
flan-t5 *	0.445	0.180	0.428	0.693
Llama-2-7b-hf *	0.348	0.124	0.332	0.637
Llama-2-7b-chat-hf *	0.408	0.126	0.397	0.659

**Table 7: Extractive QA manual annotation. Each answer is labelled between 0-2 by expert annotators.**

	Correct	Complete
t5-base	1.16	1.02
unifiedqa-t5-base	<b>1.46</b>	<b>1.32</b>
flan-t5-base	1.16	1.16
Llama-2-7b-hf	1.32	1.22
Llama-2-7b-chat-hf	1.38	1.18

1: somewhat, 2: fully). In contradiction to the metrics, annotators agree that the generative models and UnifiedQA perform better than the baseline, with UnifiedQA answers ranked significantly better (Table 7).

We investigate why metrics penalise generative QA output. We notice that generative models are more likely to ignore the prompt asking for an extractive answer and hallucinate the output format. The metrics can't capture this - the extractive token-wise metrics penalise any output outwith the original context. If a generative model rewrites the span or adds explanations for model responses, the model's metric score decreases.

Next, we review model performance according to the question type taxonomy. Model performance on causal, complex, and confirmation questions is low across models. However, the pre-trained T5 models outperform the pre-trained Llama models for factoid QA. With closer analysis, the model's tendency to add explanation penalises their metric score and causes incorrect chain-of-thought explanations (e.g. Fig. 8). Compared to this, UnifiedQA achieves an F1 score of 0.524 (Llama 2 chat: 0.317). For listing questions, generative models outperform T5 models due to T5 generating fewer tokens (F1 Llama 2 chat: 0.541 vs FLAN T5: 0.366).

In comparison, history and navigation questions require reasoning and extraction of information from previous or future steps. No model can do this well currently. T5 models outperform Llama in navigational questions for token-wise F1 (FLAN T5: 0.38 vs Llama 2 chat: 0.23). Looking at individual outputs, Llama 2 answers are often not fully wrong, but answer ambiguous questions differently to the teacher or do not follow the intended and pre-trained output structure.

### 5.3 Task Adaptation evaluation

GRILLBot modifies the task for users by replacing ingredients replacement or adopting to dietary restrictions. Replacement generator and task rewriter input and output structured data. The generative component follows the structured format end-to-end in 60% of cases. In the live conversations with correctly formatted generations, the task rewriter rewrote the task for 56% of replaced ingredients generated by the replacement generator correctly. Overall, users accepted 34% of the suggestions given by the deployed system. We evaluate whether the suggestions recommended by the LLM were factually correct and would work in practice by reviewing 25 conversations with accepted changes. According to our annotations, 73% of replacements would work, 18% were not common and 9% were incorrect.

To gain a better understanding of why users do not accept a suggested replacement, we hand-annotate conversations. We randomly sample 50 unaccepted replacement suggestions and categorise user behaviour. Most users do not accept the replacement since they ask for a new replacement suggestion, for other reasons including starting a new search or continuing with the original task.

## 6 CONCLUSION

In this work, we tackle the challenge of effective and efficient use of LLMs in interactive multimodal assistants. We decompose the task into submodules and Discussing tradeoffs in latency, correctness and fluency, we show that a hybrid approach using LLMs and specialised models for different components enables a fluent, knowledgeable, and dynamic assistant. GRILLBot helps users overcome challenges as the task processes in the real world - possibly in new and unexpected ways. For reproducibility, we continue to publish key components of GRILLBot as part of the OAT framework to allow quick deployment of similar assistants for the community. In addition, we release a new task-oriented complex QA dataset *WoTe*.

Constraints in using LLMs live are response times and computational resources needed, which is why we still often use smaller-scale specialised models with lower accuracy. However, with model distillation, we can deploy higher-quality models with lower latency. Using distilled models, we can perform model chaining where a model's output is the input for a larger model using a specific routing framework. Second, a drawback of the generation abilities of LLMs is hallucinations. In our system, generative models hallucinate system abilities and unrealistic tasks and generate potentially dangerous responses. Therefore, another line of work is to create specialised models that guardrail inputs and outputs to generative models and enforce model grounding to build even more complex pipelines with more LLMs in the loop (e.g. [13, 29]).

## ACKNOWLEDGMENTS

The authors would like to thank the rest of the GRILLBot team and the GRILL lab for their ongoing support, specifically Andrew Ramsay. This work is supported by the Amazon Alexa Prize TaskBot Challenge. It was also supported by the Engineering and Physical Sciences Research Council grant EP/V025708/1.



## REFERENCES

- [1] Eugene Agichtein, Michael Johnston, Anna Gottardi, Cris Flagg, Lavina Vaz, Hangjie Shi, Desheng Zhang, Leslie Ball, Shaohua Liu, Luke Dai, et al. 2023. Alexa, let's work together: Introducing the second alexa prize taskbot challenge. *2nd Proceedings of the Alexa Prize Taskbot Challenge 2* (2023).
- [2] Amazon. 2023. Amazon Skill Kit. <https://developer.amazon.com/en-US/alexa/alexa-skills-kit>
- [3] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz—a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), 5016–5026.
- [4] Giovanni Campagna, Silei Xu, Mehrad Moradshahi, Richard Socher, and Monica S Lam. 2019. Genie: A generator of natural language semantic parsers for virtual assistant commands. In *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. 394–410.
- [5] Jason Ingyu Choi, Saar Kuzi, Nikhita Vedula, Jie Zhao, Giuseppe Castellucci, Marcus Collins, Shervin Malmasi, Oleg Rokhlenko, and Eugene Agichtein. 2022. Wizard of tasks: A novel conversational dataset for solving real-world tasks in conversational settings. In *Proceedings of the 29th International Conference on Computational Linguistics*. 3514–3529.
- [6] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).
- [7] Rafael Ferreira, Diogo Tavares, Diogo Silva, Rodrigo Valério, João Bordalo, Inês Simões, Vasco Ramos, David Semedo, and Joao Magalhaes. 2023. TWIZ: The wizard of multimodal conversational-stimulus. In *Alexa Prize TaskBot Challenge 2 Proceedings*.
- [8] Sophie Fischer, Carlos Gemmell, Iain Mackie, and Jeffrey Dalton. 2022. VILT: Video Instructions Linking for Complex Tasks. In *Proceedings of the 2nd International Workshop on Interactive Multimedia Retrieval*. 41–47.
- [9] Sophie Fischer, Niklas Tecklenburg, Philip Zuber, Eva Kupcova, Ekaterina Terzieva, Daniel Armstrong, Carlos Gemmell, Iain Mackie, Federico Rossetto, and Jeff Dalton. 2023. GRILLBot-v2: Generative Models for Multi-Modal Task-Oriented Assistance. *2nd Proceedings of the Alexa Prize Taskbot Challenge* (2023).
- [10] Carlos Gemmell, Sophie Fischer, Iain Mackie, Paul Owoicho, Federico Rossetto, and Jeff Dalton. 2022. GRILLBot: A flexible conversational agent for solving complex real-world tasks. *1st Proceedings of the Alexa Prize Taskbot Challenge* (2022).
- [11] Carlos Gemmell, Sophie Fischer, Federico Rossetto, Paul Ochiwo, Iain Mackie, Philip Zuber, Niklas Tecklenburg, and Andrew Ramsay. 2023. Open Assistant Toolkit [OAT]: A research Platform for Multi-Modal Task Oriented Agents. <https://github.com/grill-lab/OAT>.
- [12] Google. 2023. Dialogflow. <https://cloud.google.com/dialogflow>
- [13] Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674* (2023).
- [14] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.
- [15] Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020), 1896–1907.
- [16] Chandra Khatri, Behnam Hedayatnia, Anu Venkatesh, Jeff Nunn, Yi Pan, Qing Liu, Han Song, Anna Gottardi, Sanjeev Kwatra, Sanju Pancholi, et al. 2018. Advancing the state of the art in open domain dialog systems through the alexa prize. *2nd Proceedings of the Alexa Prize SocialBot Grand Challenge 2* (2018).
- [17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [18] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2018. Recipe1M+: a dataset for learning cross-modal embeddings for cooking recipes and food images. *arXiv preprint arXiv:1810.06553* (2018).
- [19] Andrea Morales-Garzón, Juan Gómez-Romero, and Maria J Martin-Bautista. 2021. A word embedding-based method for unsupervised adaptation of cooking recipes. *IEEE Access* 9 (2021), 27389–27404.
- [20] OpenAI. 2022. Chatgpt: Optimizing language models for dialogue. <https://openai.com/blog/chatgpt/>
- [21] Paul Owoicho, Ivan Sekulic, Mohammad Aliannejadi, Jeffrey Dalton, and Fabio Crestani. 2023. Exploiting simulated user feedback for conversational search: Ranking, rewriting, and beyond. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 632–642.
- [22] Ashwin Paranjape, Abigail See, Kathleen Kenealy, Haojun Li, Amelia Hardy, Peng Qi, Kaushik Ram Sadagopan, Nguyet Minh Phu, Dilara Soylu, and Christopher D Manning. 2020. Neural generation meets real people: Towards emotionally engaging mixed-initiative conversations. *3rd Proceedings of the Alexa Prize SocialBot Grand Challenge 3* (2020).
- [23] Chantal Pellegrini, Ege Özsoy, Monika Wintergerst, and Georg Groh. 2021. Exploiting Food Embeddings for Ingredient Substitution.. In *HEALTHINF*. 67–77.
- [24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
- [25] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (2016), 2383–2392.
- [26] RASA. 2023. RASA. <https://rasa.com/>
- [27] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <https://arxiv.org/abs/1908.10084>
- [28] Alexandra Rese and Pauline Tränkner. 2024. Perceived conversational ability of task-based chatbots—Which conversational elements influence the success of text-based dialogues? *International Journal of Information Management* 74 (2024), 102699.
- [29] Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2387–2413.
- [30] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188* (2022).
- [31] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following LLaMA model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca).
- [32] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Llama: Language models for dialog applications. *arXiv preprint arXiv:2201.08239* (2022).
- [33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [34] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2023. A survey on large language model based autonomous agents. *arXiv preprint arXiv:2308.11432* (2023).
- [35] Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems* 64, 12 (2022), 3151–3195.
- [36] Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval* 17, 3-4 (2023), 244–456.
- [37] Tianyi Zhang\*, Varsha Kishore\*, Felix Wu\*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.
- [38] Diliara Zharikova, Daniel Kornev, Fedor Ignatov, Maxim Talimanchuk, Dmitry Evseev, Kseniya Petukhova, Veronika Smilga, Dmitry Karpov, Yana Shishkina, Dmitry Kosenko, et al. 2023. DeepPavlov dream: platform for building generative AI assistants. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 599–607.

U: Do you have Spanish recipes  
 S: Certainly! How about these three matches? they look so tasty! First is: gluten-free flour tortillas recipe by ...  
 U: Can you do that with a Spanish accent ...  
 S: Yes, of course! I can do that with a Spanish accent. Just wanted to say, you can start a new search by saying "cancel" or "restart".  
 U: New recipes  
 S: Certainly! How about these three matches? They look so tasty! First is: Spanish-style padron peppers, second: empanadas ...  
 U: What is inside of empanadas in Spain ...  
 S: In Spain, the most popular ingredients in empanadas are usually beef, chicken, potatoes, onions, garlic, and spices.

**Figure 7: Selected reworded example conversations of live users with fluent and grounded GRILLBot.**

Q: How long are the garlic and onions supposed to cook for?  
 Golden: cook until golden  
 Llama-2-7b-hf: 5 minutes  
 Llama-2-7b-chat-hf: According to the context, the garlic and onions are supposed to cook for 5 minutes, as stated in the following snippet: "Add onions and garlic to skillet and cook until golden."

**Figure 8: Llama-2 hallucinations for *food-59-3*. "5 minutes" are never mentioned in the task data (linked here)**

User: turn on the light  
 >> turn\_on()  
 System: I'm sorry, I'm not able to turn on or off the lights. But anyway if you don't know what to search for [...]  
 User: play some smooth jazz  
 >> play\_music()  
 System: Sure, I can play some smooth jazz. What kind of jazz do you want to hear?  
 System: If you would like, I can change the recipe to use what you have at home. Just ask me to replace an ingredient. [...]  
 User: place an ingredient  
 >> ask\_me()  
 System: Funny enough, I just wasn't sure what to do with what you said. [...]

**Figure 9: Example of NDP hallucinations handled by the LLM Fallback prompt. » signals NDP code generation.**

**Figure 5: Prompt fed into the Alpaca model when the NDP generates a system action that doesn't have built back-end logic, aka no system action should be performed live.**

```
### Instruction: You are a friendly AI assistant who is assisting a human. Respond to the human or ask a question back. Try to not repeat what you said previously. You specialise in cooking, arts & crafts, and DIY. You do not reveal your name in the spirit of fair competition. You cannot play music, games or quizzes. You are not able to read the news, turn on a light, or give recommendations for things outside cooking and DIY domains.
### Input:
You: {last_system_response}
Human: {user_utterance}
### Response: Your response:
```

**Figure 6: Prompt fed into the Llama 2 7b for QA.**

```
<s> [INST] «SYS» You are a friendly assistant who helps people with cooking and DIY tasks. «/SYS»
Provide a short answer to the user question from the context. If the context doesn't contain the answer truthfully say <unknown>. You are only allowed to use information from the context in your answer or say 'You are done' if the user has reached the last step.
Context: {Description} {Steps} {Ingredients}
Question: {Question} [/INST] Answer:
```

## A LLM GENERATION

Our LLM-based components generate interesting output across the system. We use hybrid approaches to constrain generation to ensure task safety and factualness based on task context and the LLM's world knowledge.

Fig. 5 shows the Fallback Prompt to the LLM, called when the NDP generates a system action that is not in the pre-defined action space, i.e. there is no back-end logic existing to translate this action into a system call. This means this is called when the action is beyond system capabilities. Fig. 9 shows examples of this. We show selected action codes that the Fallback handles, with various success. Despite our best attempts to constrain hallucinations, sometimes the LLM still manages to hallucinate system abilities such as playing music. The dangerous classifier does not flag this as dangerous, and the LLM did not follow the prompt passed in.

Title :  
cucumber , radish and seaweed salad

| Description :  
noodlelike black seaweed strands make this  
strikingly colorful salad a healthful side  
dish for pairing with fish , grilled tofu or  
noodle dishes . the salad benefits from at  
least 30 minutes in the refrigerator to  
marinate in the vinaigrette .

| Ingredients :  
1 cup (1/2 ounce) dried arame seaweed  
2 large cucumbers , halved lengthwise and thinly sliced  
1 bunch (about 8) small red radishes , trimmed and  
quartered  
2 tablespoons unseasoned rice vinegar  
2 teaspoons reduced-sodium tamari  
2 tablespoons black or white sesame seeds , toasted  
and cooled (optional)

| Steps :  
soak arame in cold water until tender , about 15  
minutes . ;  
drain and transfer to a large bowl . ;  
add cucumbers , radishes , rice vinegar and tamari and toss to combine . ;  
cover and chill for at least 30 minutes .  
just before serving , toss vegetables together  
again and sprinkle with sesame seeds .

Figure 10: Examples of QA context and user questions asked about Wizard-of-Task-food-135. Colourful text corresponds to the annotation of what span answers the each coloured question.

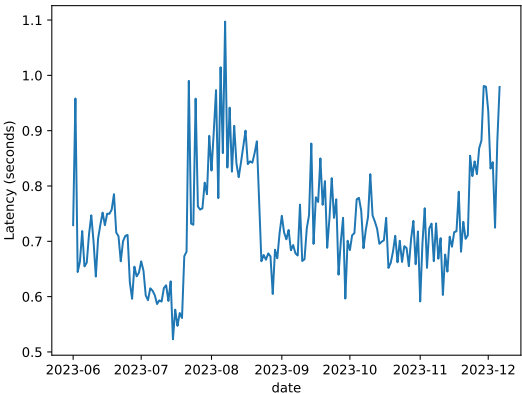


Figure 11: Average Latency since GRILLBot v2 went live

Table 8: Rejected Replacement Analysis

Reason	Amount
New Search	18%
Ignored Replacement	18%
Another Replacement Request	38%
Exit	12%
System parsing error	14%

Table 9: Generated ingredient replacements by the task adaptation component.

Original	Replacement	Recipe title
eggs	eggs substitute	Rice Pudding
dried ginger	fresh ginger	Drunken Chicken Recipe
sundried tomatoes	fresh tomatoes	Mediterranean Chicken
dried apricots	fresh apricots	Mediterranean Chicken
peanut oil	olive oil	Firecracker Grilled Salmon
baking powder	baking soda	Pumpkin Bread
fengryk seeds	ground cumin	Cook in Curry Sauce
sauerkraut	pickled cabbage	Tenderloin Sandwiches
milk	whipped cream	Spaghetti & Meatballs
thai apple eggplants	regular eggplants	Gaeng Om Gai
pancetta	bacon	Christmas Pasta
pecorino romano	parmesan	Spaghetti Carbonara
black pepper	cayenne pepper	Grilled Chicken Breasts

B TASK ADAPTATION

Table 9 shows a few generations of how individual ingredients could be replaced as asked for by the user. Table 8 shows hand-annotated reasons why the user rejected the replacement the system requested.

C WOTE CREATION

Fig. 10 shows an example of an annotated question’s task context during the creation of WoTe. The blue extract corresponds to the annotated response for "Is the vinaigrette part of the recipe or should I be using a store-bought bottle?" [food-135-1] and the green text corresponds to the response to "I think that it looks really yummy, and your response doesn’t tell me about the cucumber. How much cucumber will I use in this dish?" [food-135-4].

D LATENCY OF COMPONENTS

A challenge we balance throughout the competition is that larger models tend to be more fluent and knowledgeable, but result in higher response latency. Additionally, when the user load increases, this can slow down the inference time of models. We constantly balance between improving the system’s abilities and maintaining low latency. Fig. 11 shows the average latency for end-to-end responses in our system. In Calendar Week 29, we deployed more computationally heavy components, resulting in a higher average latency but more fluent responses.