NEWS LETTER

24

June | 2026

CLOUD STARS

Dublin, Rehearsal of Castellers d'Éire, Portobello Community Center

Large Language Models Fine Tuning with Differential Privacy

by Ferran Agulló

(Barcelona Supercomputing Center - Universitat Politècnica de Catalunya)

Following the successful collaboration between the Barcelona Supercomputing Center and IBM T.J. Watson Research Center in New York, which included two secondments by Pol Garcia Recasens and Ferran Agulló, also a collaboration from Joan Oliveras, resulting in three publications (Recasens et al., 2024; Recasens et al., 2025; Agulló et al., 2025), the collaboration has now expanded to IBM Research in Dublin, Ireland. This new phase started with a secondment by Pol Garcia last year and continues with the current one by Ferran Agulló.

While the collaboration with IBM T.J. Watson focused on optimizing LLM inference through low-level analysis, memory management, and scheduling, the work with IBM Research Ireland has shifted toward privacy in LLMs. In 2025, Pol Garcia Recasens, among other topics, investigated how malicious contributors can inject bias into synthetic datasets, resulting in a workshop publication and an ongoing revision for a top-tier journal. Currently, Ferran Agulló is undertaking his secondment until the end of July, where he works on differentially private fine-tuning of LLMs.

Differential privacy (DP) has gained significant attention in recent years because large language models tend to memorize aspects of their training data. Although this behavior is not always harmful, it can pose serious risks in sensitive domains such as healthcare, where LLMs can be highly beneficial across many applications. Training these models requires access to data, and this process can lead to unintended memorization. As a consequence, external users may extract information about the training dataset, even when the data itself is not public and only the trained model is available. These risks include membership inference attacks, where adversaries determine whether specific samples were part of the training set, as well as more severe threats such as data extraction and the leakage of personally identifiable information, including phone numbers or email addresses.

The goal of the secondment is not only to build expertise in this area and to perform exploratory analysis and engineering work for specific use cases, but also to contribute to the advancement of the field through a publication in a top-tier venue after the secondment concludes.

Beyond research activities, Dublin offers many opportunities for personal growth. Although it is a relatively small city, it is vibrant and diverse, with a rich mix of cultures that encourages interaction and exchange. One example is

Castellers d’Éire, shown above, a traditional Catalan practice that has also taken root in Dublin. This activity, based on building human towers, brings together Catalan, Irish, and international participants, creating a shared environment where people collaborate and have fun.

cloudstars.eu | twitter.com/Cloudstars_2023 | github.com/cloudstars-eu

CLOUDSTARS project has received funding from the European Union's Horizon research and innovation programme under grant agreement No 101086248