TECHNOLOGICAL UNIVERSITY DUBLIN

MASTERS THESIS

Enhancing Static Malware Analysis with Large Language Models and Retrieval-Augmented Generation

---

Author:
Andre M M FARIA

Supervisor:
Dr Robert G SMITH

A thesis submitted in fulfillment of the requirements
for the degree of M.Sc
in Applied Cybersecurity

in the

School of Informatics and Cyber Security

May 2025

Declaration of Authorship

I, Andre M M FARIA, declare that this thesis titled, "Enhancing Static Malware Analysis with Large Language Models and Retrieval-Augmented Generation" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this Institute of Technology Blanchardstown.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: March 5, 2025

"The true masters are the friends we make along the way."

Anonymous

TECHNOLOGICAL UNIVERSITY DUBLIN

Abstract

School of Informatics and Cyber Security

M.Sc

Enhancing Static Malware Analysis with Large Language Models and Retrieval-Augmented Generation

by Andre M M FARIA

Malware's quick growth presents serious cybersecurity concerns, necessitating ongoing innovation in methods for detection, analysis, and mitigation. When it comes to handling complex and adaptable threats, traditional malware analysis techniques, which mostly rely on manual labour and rule-based systems, are progressively less effective. Large language models (LLMs) are a revolutionary way to automate malware analysis, even though Artificial Intelligence (AI) approaches have been effectively incorporated into cybersecurity. Current AI solutions, including machine learning classifiers and clustering algorithms, concentrate on aspects of malware but frequently lack the overall skills necessary to manage intricate datasets or fully understand virus behaviors.

This study suggests using LLMs to improve and automate static malware analysis. It seeks to automate crucial components of malware reverse engineering, such as code analysis, possible behavior interpretation and anomaly detection, by leveraging LLMs' capacity to handle and synthesize massive, heterogeneous information. In addition to addressing the drawbacks of manual and conventional AI techniques, this strategy adds scalability and adaptability to quickly changing malware threats.

In order to optimize the efficacy of malware analysis in cybersecurity, this study proposes a methodology that combines domain-specific datasets, such as malware sample databases, with prompt engineering techniques to evaluate samples and identify parameters such as malware classification categories and behavioral patterns on decompiled code. This approach is expected to provide a more comprehensive and accurate analysis of malware samples, as well as to improve the overall efficiency of malware analysis in cybersecurity.

Keywords: Malware Analysis, Large Language Models, Static Analysis, Cybersecurity, Artificial Intelligence, Generative Artificial Intelligence, Machine Learning, Reverse Engineering, Anomaly Detection, Code Analysis.

Acknowledgements

Thanks for my parents that supported all of my bullshit over the years…

Thanks to my supervisor, Dr Robert G Smith, for his guidance and support. :D

Contents

List of Figures

List of Tables

List of Abbreviations

LLM **L**arge **L**anguage **M**odel(s)
RAG **R**etrieval-**A**ugmented **G**eneration

Chapter 1

Introduction

This is the introduction of the thesis.

1.1 Background

Ever since the creation of the first computer by Charles Babbage in the 19th century, computers have been evolving at a rapid pace performing more and more tasks as the time passes. With this evolution, some people began to notice that manners in which to subvert the established systems. These individuals, by force of belief or personal gain, have caused losses in the trillions to organizations and individuals across the globe.

As a response to these threats, these organizations and/or individuals began to develop techniques and tools to counter malicious actions. It has then, started the cat-and-mouse race between criminals and agents of the law. One of these techniques revolves around the analysis of what software adversaries develop to understand how they operate and the vulnerabilities they exploit to weaken or outright topple established security systems.

As a basis, there are two types of analysis of software that can be made to understand any kind of software. These are as follows:

- **Static Analysis**: Focuses on looking at the sequences of individual instructions on the software bytecode to identify the its characteristics, such as identification of the parameters in which the analyzed software is built upon (e.g. Operating system it executes on, processor architecture, etc), what execution paths it employs, what system calls it performs, etc. This analysis is done without running the code and it requires a lot of time and effort from the analyst.

- **Dynamic Analysis**: Entails running the malware to gather runtime information such as network calls and system call execution data. This analysis is done by creating a special environment to have the analyzed software to run and gather the mentioned data.

The

1.2 Definitions

Here are defined all of the concepts and/or tools that are mentioned on this document.

1.2.1 Large Language Models

Large Language Models (LLMs) are a type of multi-parameter Machine Learning (ML) algorithm designed for natural language processing. These models that are trained on a large volume of text using a multitude of learning techniques such as self-supervised learning.

Generative pre-trained transformers (GPTs) are the biggest and most powerful LLMs. Current models can be adapted to certain tasks (via fine-tuning) or directed by prompt engineering [Brown et al., 2020]. These models learn to predict the syntax, semantics and ontologies found in human language, but they also pick up biases and errors from the training data [Manning, 2022]. Popular examples of LLMs include:

- GPT models (like GPT-4 or ChatGPT)
- Google's PaLM and Gemini
- Meta Llama series models

1.2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a technique that allows for LLMs to query information from a specific type of datastore and use it as context for the response it provides. This allows for adaptive learning of the models without the need of retraining (fine-tuning). It improves the quality of LLM output as the model does not rely solely on pre-trained knowledge but also on relevant information from the datastore.

RAG works by having the LLM to query the datastore using vector search, embeddings or keyword matching (depending on the datastore type), feeding this data onto the LLM input and then continuing with the normal LLM process for output generation.

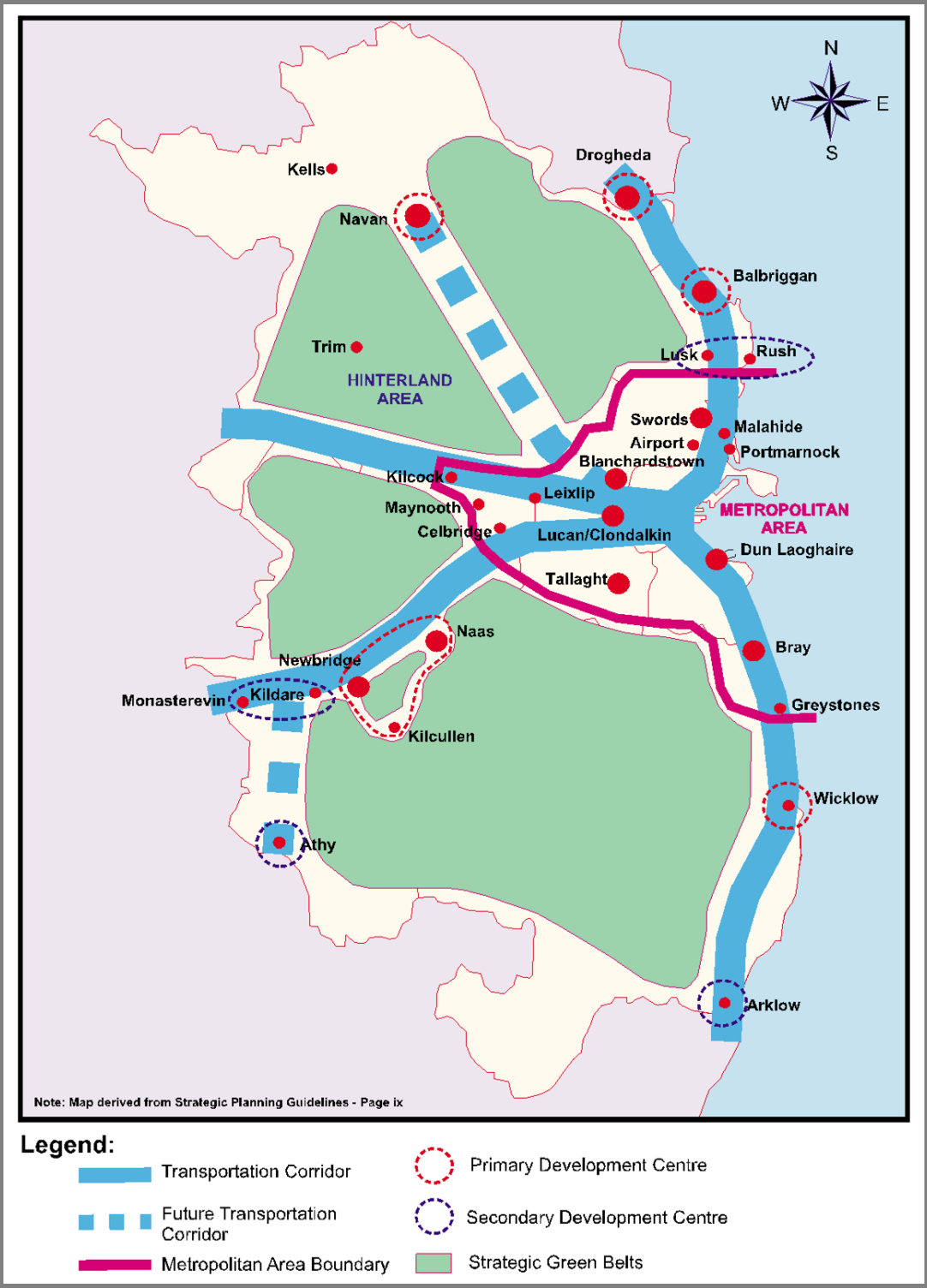
1.2.3 Malware

Short for *Malicious Software*, Malware is any program or code that is created for malicious purposes like exploitation of computer systems networks or users. Often created with monetary gain in mind, these programs activities include stealing sensitive data, damage system, disrupt operations or gain unauthorized access to systems or environments. Common types of malware include:

- Viruses: Attach themselves to legitimate software, spreading when the software is run.
- Worms: Spread automatically over networks, replicating rapidly without user intervention.
- Trojans: Disguised as legitimate software, tricking users into installing them to enable unauthorized access.
- Ransomware: Encrypts or locks files, demanding payment (ransom) for restoration.
- Spyware: Secretly collects sensitive information like passwords or browsing activity.
- Adware: Delivers unwanted advertisements, potentially slowing down or compromising systems.

1.2.4 Malware analysis

1.2.5 Static Malware Analysis



1.3 Research questions and objectives

1.3.1 Questions

These are the research questions that will guide the development of this thesis.

- How can LLMs be leveraged to enhance static malware analysis?
- How does a RAG-enhanced LLM compare to traditional static analysis techniques in terms of accuracy, efficiency, and interpretability?
- How does RAG improve the accuracy and explainability of malware classification using static features?
- Can a RAG-based LLM framework provide actionable threat intelligence insights from external malware databases?

1.3.2 Objectives

- Develop a methodology for integrating LLMs with RAG to enhance static malware analysis.
- Evaluate the effectiveness of RAG-enhanced LLMs in identifying and explaining malware threats using static features (e.g., file structure, bytecode, API calls).
- Compare the proposed RAG-enhanced LLM-based approach against traditional static malware analysis techniques in terms of accuracy, efficiency, and interpretability identifying challenges and potential risks (e.g. adversarial manipulation and hallucination).

Chapter 2

Literature Review

2.1 What is a Literature Review?

A literature review is a section of your thesis or dissertation in which you discuss previous research on your subject. Following your Introduction chapter, your literature review begins as you try to answer your larger research question: Who has looked at what, why, and what have they found? It allows you to understand what others have said about your topic, to verify your assumptions, to refine your initial research question, and to identify gaps. For your readers, the literature review also demonstrates that you are knowledgeable about related research and scholarly traditions in your field.

2.1.1 Preparing to Write

The literature review is more than just a list of previous research papers in the field. If you think of writing a thesis or dissertation as writing a story of your research, the literature review then will be a story within a story. In the literature review story, you tell the reader about general trends, traditions, and approaches to your subject, ones that surround and support your study.

Choose texts to help you try to answer your research question. As you explore the literature, take notes:

- Why did you pick up this text? [Reminder: What is being studied, by whom, why? What did they find? As you pick up a text, note all documentation information.]
- How does this article, chapter, book, study help you answer your question or not?
- When you find a publication of interest, read the Abstract to see if it is what you are looking for. If not, discard the text. If it does seem to be what you are looking for, then glance over the Introduction and Conclusions. Again, if it is what you are looking for, you can now invest the time to read the entire publication, or the section of the publication that interests you. This method save a lot of time in the long term.
- Be sure that all publications are from a credible source: You can gauge this by where an article/book has been published, if it has been peer reviewed, how it has ben written, how many times it has been cited etc...

After you have read and written, draw a diagram, chart, or matrix that would help you to visualize connections between your sources and reveal a possible structure for your literature review. Some researcher like to print papers and organise

content with colour (with highlighters and post-it notes), others like to use tools such as *NVivo*. This approach allows you to notice distinct patterns in the literature, e.g., how an algorithm has developed over time. You may choose to plot it out on a timeline. Or, you may decide to organize your literature review by the researchers' stance towards your subject. Or, you may want to create a sort of bubble map to discover:

- What major trends and patterns in the results of previous studies emerge?
- What common threads do you find?
- How do these studies connect?

There is no right or wrong way for structuring the review. It should explain the thinking process behind your choices and help reveal the need to answer your question (to fill a gap) and how to go about doing that (the methodology).

When you have a rough draft completed, ask yourself:

- What previous research has been more significant and less significant?
- What gaps in literature have you noticed? Why do these gaps exist?
- How might your research hypothesis or research questions inform your organization and characterization of the previous literature?

2.1.2 Revising

When describing, critiquing, and citing your sources, use the following citation patterns to introduce and comment on sources:

- Generalisation (combining 2 or more sources): Describe what makes this group of sources a category
- Summarise each key source; paraphrase the author[s]' argument (this is not plagiarism because you are citing the work).
- Try to avoid using quotations to note key words or phrases... better not to overuse this strategy and to use your own words where possible.
- Use block quotations (more than 40 words) sparingly.

However, avoid ambiguous citations like these two:

In the example above, it is not clear whether Clement and Lee are major researchers in their fields or what their work includes. Also, one author does not suggest "wide investigation" or "much" research. Best to use multiple sources for broad statements like these.

Help your readers make their way through your literature review by referring to its organization or back to a part of the review, or by providing a definition. For example, use words and phrases, such as *In this section, I will discuss ...*
This part will describe ...
For the purpose of this discussion, metadiscourse means ...
The main purpose of this review has been ...
Thus far, this review has outlined ...

Things to Remember

- Avoid describing each piece of relevant research in detail, piece by piece.
- Focus on general trends and approaches.
- Only critique the few most relevant, seminal sources. There is no need to critique each source.
- When reviewing a study, avoid reporting an author's assertions as though they were findings.
- Highlight agreement before disagreement.
- Depending on your field of study, you may want to tell a story that led you to this research and would help explain your choices to include or exclude previous research.

2.1.3 Sources

They Say/I Say: The Moves That Matter in Academic Writing by Gerald Graff and Cathy Birkenstein Academic Writing for Graduate Students and Telling a Research Story: Writing a Literature Review by Christine B. Feak and John M Swales
https://www.jsums.edu/wrightcenter/files/2016/03/Writing-a-Literature-Review.pdf

2.1.4 Citing and Referencing

This Latex template uses the 'natbib' package to manage references and citations. There is a good introduction to this package here: https://www.overleaf.com/learn/latex/Bibliography_management_with_natbib

Note: there are different referencing styles, these can be set in the main thesis.tex file. Find out which style you should use from your project documentation, project coordinator or supervisor.

It is important to note that when referencing in-text, you should format the citation differently depending on how you reference the author. Consider the following sentence:

"*Smith, 2023 explores the use of Association Rules Mining to identify patterns in a sign language dataset.*"

You will note that when Smith's 2023 publication is referenced directly in the text, only the year of the publication is in brackets, i.e., the authors name is not in brackets.

"*The use of Association Rules Mining to identify patterns in a sign language dataset has been explored recently (Smith, 2023) .*"

When the same publication is referenced indirectly at the end of the sentence, the authors name and year of publication are inside the brackets. See the following reference sheet to help you keep track of this: https://gking.harvard.edu/files/natnotes2.pdf

Chapter 3

Methodology

3.1 What is a Methodology?

Every thesis, regardless of the discipline and field of inquiry it relates to, needs to answer these questions:

- How did you do your research?
- Why did you do it that way?

This covers not only the methods used to collect and analyse data, but also the theoretical framework that informs both the choice of methods and the approach to interpreting the data. In some disciplines, the approach to knowledge underpinning both the type of research questions asked and the methods chosen to answer them is called "methodology", and needs to be articulated. Both methods and theoretical approach relate explicitly to the research question(s) addressed in the thesis.

You may need to summarise available methods and theoretical approaches for your research topic; you will certainly need to justify your choice of method(s). If you apply a combination of methods you'll need to justify why you chose such an approach. Your explanation should also indicate any reliability or validity issues concerning the data, and discuss any ethical considerations that arise from your choices.

Whilst patterns of organisation in a methods chapter may vary, there are some common elements that you'll need to include to achieve an informative chapter. Let's identify these features:

- place or setting of the research
- duration of the study and other time related factors
- study design – e.g. an outline of the research stages including instruments and techniques
- specifics of the participants, materials, etc.
- sampling frameworks (e.g. criteria, size, scope, etc.)
- any inclusions/exclusions
- outcome measurement procedures (e.g. statistical tests, comparisons, etc.)
- consent and ethics committee approval
- theoretical basis of the research

- data management

While most of these elements will be relevant to your methods chapter, you'll find that there are discipline specific elements and requirements. The detail and emphasis of what is covered in a discussion of methods/methodology will be different in different disciplines.

3.1.1 STEM specific Method Chapter

Key features of method descriptions in STEM disciplines include:

- demonstration of fit between methods chosen and research question(s)
- rationale for choosing materials, methods and procedures
- details of materials, equipment and procedures that will allow others to:
  - replicate experiments
  - understand and implement technical solutions

Chapter 4

Discussion of Results

4.1 Section Introduction

The reporting and discussion thesis chapters deal with the central part of the thesis. This is where you present the data that forms the basis of your investigation, shaped by the way you have interpreted it and developed your argument or theories about it. In other words, you tell your readers the research story that has emerged from your findings. These chapters will form the bulk of your complete thesis. Before you even begin writing up the reporting and discussion chapters, you'll need to undertake some thinking and planning.

There is quite a loot to say about this topic so Ive provided a like here for further reading: https://www.monash.edu/student-academic-success/excel-at-writing/how-to-write/thesis-chapter/reporting-and-discussion-thesis-chapters

Chapter 5

Conclusion

5.1 About Conclusions

Depending on the type of research presented in the thesis, conclusion chapters or sections tend to include at least some of the following:

- A clear answer to your research question or hypothesis
- Summary of the main findings or argument
- Connections between your findings or argument to other research
- Explanation and significance of the findings
- Implications of the findings
- Limitations of the research and methodology
- Recommendations for future research

Your conclusion chapter is the place to emphasise the new knowledge that you've contributed to the field of study and explain its significance. This chapter is your opportunity to leave a strong impression on the reader (assessors) about the strength and relevance of your research, and your skills as a researcher.

Importantly, the conclusion chapter must link with your introduction chapter to complete the framing of the thesis and demonstrate that you have achieved what you set out to do.

Appendix A

Frequently Asked Questions

A.1 Getting Feedback

1. Get feedback **often** and from different audiences – your family, friends, professors, colleagues, advisor, other graduate students. The more you talk about your research, the more comfortable you get with it.

2. Keep a positive attitude. Research is hard. If it were easy, everyone would be doing it.

3. Consider setting up or joining a thesis group to share your ideas and experiences.

Supervisor's feedback

Some supervisors will ask for you to send each chapter as you complete it, offering feedback at that point, and then again at the end when the thesis chapters are collated. Other supervisors may want to be more involved, and there are others who will not want to see your thesis until it is completed by your standards. Whichever approach your supervisor takes, be aware that they will need some time to read through your work and provide feedback. Your thesis review is likely not the only piece of work your supervisor is undertaking, so be patient and factor review time into your work schedule.

A couple of points to note about supervisor feedback:

- You will receive feedback on your approach to research (i.e., method, experiment design etc...) as well as your writing. It is your responsibility to take notes at meetings etc. in order to record this feedback. It is also up to you whether or not you act upon the feedback provided.

- Your supervisor's role is to guide your work. It is not their job to complete the research, suggest methods, design experiments, or to write/rewrite sections of your thesis.

- Your supervisor should be supportive but sometimes their feedback may be difficult to hear. Just remember, their goal is to guide you and to make you a better researcher. Learn to have your work criticised in a constructive manner, it is part of the learning process.

- It is not the role of your supervisor to proofread your thesis. Many supervisors will point out typos, grammatical errors or styling issues etc. when they see them, but this is not their role.

A.2 Proofreading/copyediting

It is important to have your work proofread[1]. If English is not your first language, this is even more important for you.

How?

A good approach is to proofread yourself as you write and then again when you are finished writing a section or chapter. When you have a near final draft, have it proofread by a friend, family member, colleague, or a classmate etc... (not your supervisor). Choose your proofreader wisely. Make sure that they have good written English skills and are able to spot grammatical errors. A native English speaker can be good for this but not all native English speakers have the skills needed to be a good proofreader.

There are many things to look out for when reviewing your own work, everything from text alignment and section numbering, to figures and tables, to spelling and grammar. It's best to identify and fix any of these errors immediately. Don't wait until the end because these will build up and it often takes longer than you think to fix them.

If you find that you make the same mistake regularly, e.g., you misspell the same word regularly, or you use a colon where you shouldn't, then make a list of these to check back when you are finished each section (the search feature is good for this).

[1] Two types of editing that are commonly used interchangeably are copy editing and proofreading. Both types of editing clean up writing, but each has its distinct contribution to the process. https://thesiswhisperer.com/2016/11/30/doing-a-copy-edit-of-your-thesis/

A.3 Writing Assistants

In the past, students may have used tools such as Grammarly or Quetext, but this has become more problematic because such editing tools now come with AI assisted writing (see more here: https://tudublin.libguides.com/c.php?g=720901&p=5233062).

Using tools such as a spell checker, a grammar checker, and a punctuation checker are generally acceptable. Using more advanced tools to rewrite sentences, check tone, offer alternative word choices, offer citations etc... is not acceptable.

If in doubt don't use any such software. In general, it appears that, as of 2025, the free version of Grammarly is fine to use, but the pro version is not.

A.4 Example of Longtable

Ticket Type ID	Description
300	Feeder Ticket - Child
301	Feeder Ticket - Adult
310	10-Journey Feeder - Adult
317	Airlink Adult Airport-Busarus
318	Airlink Child Airport-Busarus
319	Airlink Child Airport-Heuston
320	Airlink Adult Airport-Heuston
333	Adult Single Feeder
365	Child Bus/Rail Short Hop - Day
366	Adult Bus/Rail Short Hop - Day
367	Family Bus/Rail Short Hop - Day
369	4 Day Explorer
410	Weekly Adult Short Hop Bus/Rail
430	Weekly Adult Medium Hop Bus/Rail
431	Weekly Adult Long Hop Bus/Rail
432	Weekly Adult Giant Hop Bus/Rail
433	Monthly Adult Short Hop Bus/Rail
455	Monthly Adult Long Hop Bus/Rail
456	Monthly Adult Giant Hop Bus/Rail
457	Monthly Student Short Hop Bus/Rail
458	Annual Bus/Rail
478	Annual All CIE Services
479	Annual CIE Pensioner Bus/Rail
480	Monthly CIE Pensioner Bus/Rail
493	Foreign Student - 1 Week
494	Foreign Student - 2 Week
495	Foreign Student - 3 Week
496	Foreign Student - 4 Week
497	CYC Group
600	Adult Cash Fare
608	Nitelink (Maynouth/Celbridge)
609	Nitelink (Maynouth/Celbridge)
610	Child Cash Fare
620	Schoolchild Cash Fare
625	Adult (formerly Shopper)
630	Adult 10-Journey (3 Stages)
631	Adult 10-Journey (7 Stages)
632	Adult 10-Journey (12 Stages)
633	Adult 10-Journey (23 Stages)
634	Adult 10-Journey (23+ Stages)
640	Adult 2-Journey (3 Stages)
641	Adult 2-Journey (7 Stages)
642	Adult 2-Journey (12 Stages)
643	Adult 2-Journey (23 Stages)
644	Adult 2-Journey (23+ Stages)
650	Schoolchild 10-Journey
651	Scholar 10-Journey
652	Schoolchild 2-Journey
653	Scholar 2-Journey
657	Transfer 90 (or Passenger Change)
658	Adult Single Heuston-CC
660	Adult One Day Travelwide
661	Child One Day Travelwide
662	Family One Day Travelwide
665	Rambler (3 Day Bus only)
670	Weekly Adult Bus

Ticket Type ID	Description
671	Weekly Adult Cityzone
690	Weekly Student Travelwide
691	Weekly Student Cityzone
705	Monthly Adult Citizone (AerLingus.)
710	Monthly Adult Travelwide
730	Annual Adult Travelwide
760	Annual Staff Bus
790	School Pass
791	OAP Pass
800	City Tour - Adult
801	City Tour - Family
802	City Tour - Child
898	10 - Journey Test Ticket

Bibliography

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL]. URL: https://arxiv.org/abs/2005.14165.

Manning, Christopher D. (May 2022). "Human Language Understanding & Reasoning". In: *Daedalus* 151.2, pp. 127–138. ISSN: 0011-5266. DOI: 10.1162/daed_a_01905. eprint: https://direct.mit.edu/daed/article-pdf/151/2/127/2060607/daed_a_01905.pdf. URL: https://doi.org/10.1162/daed_a_01905.

Smith, Robert G. (2023). "Exploiting Association Rules Mining to Inform the Use of Non-Manual Features in Sign Language Processing". PhD thesis. Technological University Dublin, Ireland.