Aalto-yliopisto
Aalto-universitetet
Aalto University

Master's Programme in Information Networks

Privacy-Friendly Click-Through Rate Prediction in Programmatic Display Advertising

Ilia Zalesskii

Master's Thesis
2026

Aalto-yliopisto
Aalto-universitetet
Aalto University

Author Ilia Zalesskii

Title Privacy-Friendly Click-Through Rate Prediction in Programmatic Display Advertising

Degree programme Information Networks

Major Strategy and Organizational Design

Supervisor Prof. Alex Jung

Advisor Mr Alessandro Rosetti (MSc)

Collaborative partner C Wire AG

Date 31 July 2026 **Number of pages** 45 **Language** English

Abstract

The digital advertising industry's revenues reached $300 billion in 2025. This massive industry is helping brands reach consumers and accelerating innovation.

However, it has also been notoriously pervasive in regard to the privacy of online users, using various tracking technologies and data brokerage. This vast data collection has been argued to be important to make ads more relevant to users and improve the campaign KPIs.

In this thesis, I set out to show how engagement metrics can be optimized without user behavior data, achieving a modest +2% Relative Information Gain over the baseline using a two-tower late fusion architecture, with cached inference of ≈200'000 predictions per second.

This outcome brings into question the reasoning behind using personal data for ad targeting and offers both theoretical and practical evidence showing the potential of privacy-first and contextual targeting approaches.

Future research may continue exploring the application of pre-trained Transformers to solve essential problems of online advertising via contextual targeting.

Keywords contextual targeting, real-time bidding, pre-trained language models, online privacy, recommender systems, digital advertising

Preface

I want to thank my colleagues from C Wire for providing me an opportunity to explore this topic.

Generative AI has been used to aid the development of the code supporting this thesis and to generate ad texts summaries (Section 3.1.2). I take personal responsibility for all of the findings and outcomes. AI was not used to produce any of the writing.

The code repository containing training, evaluation and other details will be accessible at https://ilia.fi/bertctr.

Otaniemi, 31 July 2026

Ilia Zalesskii

Contents

Abstract 2

Preface 3

Contents 4

1 Introduction 6

2 Background 8
- 2.1 Programmatic Advertising . . . 8
  - 2.1.1 Display advertising and other types of online advertising . . . 8
  - 2.1.2 Real-Time Bidding and auctions . . . 8
  - 2.1.3 Pricing models . . . 9
  - 2.1.4 Bidding strategy optimization . . . 10
- 2.2 Tracking-enabled programmatic advertising . . . 11
  - 2.2.1 Advent of behavioral targeting . . . 11
  - 2.2.2 What is tracking . . . 12
  - 2.2.3 Value of tracking . . . 13
  - 2.2.4 Negative externalities of tracking . . . 13
  - 2.2.5 Data protection laws . . . 15
  - 2.2.6 Contextual targeting . . . 17
- 2.3 Machine Learning in Programmatic Advertising . . . 17
  - 2.3.1 Recommender systems . . . 17
  - 2.3.2 Language Models . . . 18
  - 2.3.3 CTR prediction . . . 20
  - 2.3.4 Key evaluation metrics . . . 21

3 Methods 22
- 3.1 Data . . . 22
  - 3.1.1 Dataset statistics . . . 23
  - 3.1.2 Textual data . . . 23
- 3.2 Baseline: rule-based algorithm . . . 26
- 3.3 Logistic Regression . . . 26
- 3.4 Field-aware Factorization Machine (FFM) . . . 27
- 3.5 Late fusion BERT-CTR . . . 29
  - 3.5.1 Training details . . . 30
  - 3.5.2 Experiment: dimension reduction . . . 32
  - 3.5.3 Implementation . . . 34

4 Results 35
- 4.1 Experiment: dimension reduction and node-wise summation . . . 36
- 4.2 Calibration analysis . . . 37

5 Discussion 39

6 Conclusion **40**

1 Introduction

Advertising could be considered the engine of modern capitalism. It can inform consumer decisions and dramatically scale the impact of innovations. With the rise of e-commerce in the XXI century, the ability to reach customers online and effectively target them became crucial for businesses, and online advertising became integral to the monetization of the internet (Evans, 2009).

More recently, big data, analytics and artificial intelligence became key enablers of more effective digital advertising campaigns, enabling industry revenues to swell to $300 bn annually (Interactive Advertising Bureau and PwC, 2026). Predictive models are trained to decide in real time which ads to show to which user, based on the estimation of probability of a conversion or engagement.

With the average internet user being exposed to thousands of ads daily, the advertising technology companies are collecting as many valuable data points as possible, which often involves tracking users' online browsing and purchases to infer their individual preferences and characteristics such as age and gender, often without their consent (Sivan-Sevilla et al., 2025). To create persistent user identification across sessions, advertisers deployed third-party cookies, which persist in the browser and are managed by companies independent of those that the user intentionally interacts with.

These practices are intrusive: they violate the consumers' privacy, and result in unrestricted and non-consensual personal data collection and sharing, which creates possibilities of private surveillance and data leaks. This drives regulators to enforce data protection laws (like GDPR in EU and CCPA in California), and motivates more and more consumers to adopt ad-blocking practices.

In reaction to that, Apple and Mozilla restricted usage of 3rd-party cookies in their browsers, while Google (market leader) first pledged to do so, but since then seems to have walked back their plan. Revenue is perceived to be at risk from not being able to attribute collected data to individual profiles. Despite regulatory pressure, there is limited empirical evidence and research on how to achieve competitive KPIs such as click-through rate (CTR) without using personal data.

Thus, the main question this thesis aims to tackle is "How can CTR be competitively predicted in a no-tracking environment?" The research questions I focus on are as follows:

1. Can a CTR model using only contextual signals (such as time of day and text on the webpage that the ad is shown on) beat the production-level baseline?

2. How can the textual and tabular inputs be combined in predictions, and does a pre-trained language model offer an advantage over simpler methods?

3. Can such a model satisfy practical reliability and latency constraints?

In this thesis, I first provide a background on real-time bidding and ad exchange auctions, then make an argument for contextual targeting as a viable and desirable alternative to tracking. Then, I develop a dedicated model for CTR prediction based on contextual targeting. Finally, I evaluate the developed model against other possible

approaches and different architectures, and show that it achieves acceptable KPIs and fits strict system requirements even without personal data collection.

2 Background

2.1 Programmatic Advertising

2.1.1 Display advertising and other types of online advertising

The work presented in this thesis concerns programmatic digital display advertising, which is an open ecosystem that works to serve advertisements on potentially any website on the internet, for example, in the form of banners, in a way where the advertiser with the highest bid gets to show the advertisement. It is important to understand, however, that this sector is responsible only for a subsection of all the ads users encounter on the internet. Advertisements in search engines and large social media networks have their own, closed advertisement systems, where they set the prices, record users' preferences and manage targeting and delivery on their platforms. The rest of the internet, the so-called Open Web, is either managed by direct agreements, where pricing is decided in advance with exclusive partners, or open to programmatic auctions.

At its core, programmatic advertising is an automated decentralized system that "allows organizations (predominantly retailers) to bid for the privilege to publish personalized online advertising in the right place, to the right people, at the right time" (Samuel et al., 2021). One event of an advertisement rendering on a user's screen is called an `impression`. From the supply side, web publishers auction off their inventory (impressions) in real time to advertisers, usually via a supply-side platform (SSP), which connects their website with a network of buyers. Bidding is carried out on behalf of buyers by demand-side platforms (DSP), which utilize the digital information about the impression and customer, and automatically calculate (in milliseconds) whether the space is worth having and what ad to show. Ad exchanges (ADX) connect buyer and seller side, and serve as trading desks that hold auctions.

2.1.2 Real-Time Bidding and auctions

When a user visits a page with an available ad slot, a bid request is sent to the ad exchange through SSP. Then, the ADX forwards bid requests to the DSPs. Based on available data about the user and impression, each DSP decides whether to bid and how much. Once the bids are collected, the auction is commenced, and the highest bid wins, leading to its ad being shown to the user (Ou et al., 2023). This auction-style decision process is also called header bidding, implemented e.g. using the open-source Prebid framework (Prebid.org, 2026).

The whole process often finishes within 100ms to guarantee a good user experience and is thus called real-time bidding, which also puts a strong constraint on the algorithms used by DSPs. This latency requirement includes network hops, internal auctions, ad matching, CPM estimating, fraud detection and so on, with each step competing for the latency budget. Meanwhile, publishers stop accepting bids that go over the timeout and overall encourage faster bidders. This means that a candidate CTR prediction model would ideally produce sub-millisecond responses, at scale.

Older literature on the topic reports using second-price auctions, where the bidder with the highest bid wins, with payment being the second-highest bid (Ou et al., 2023). However, current standard in major ad exchanges is first-price auctions, where the winner pays exactly what they bid. First-price auctions were initially shunned because of their inherent incentive to "shade" the estimated value of the impression and try to barely outbid the competitors instead of bidding the true value of the impression. Nevertheless, the transparency and accountability of first-price auctions proved superior. Publishers also introduced "floor prices" - the lowest accepted bidding prices - to protect revenues.

Figure 1 offers an overview of data typically available in a bid request. Based on these datapoints, advertisers decide the value of this impression to them.

Bid Request Data			
Data Categories		**Common Variables**	**Example**
User data	**Data generated by user tracking**	Identifier (ID)	User- id "123-ABC-789".
		Browsing history	The user has visited www.sports.com three times already.
		Ad recency	The user saw the ad "ABC" two minutes ago.
		Ad frequency	The user saw the ad "ABC" already four times.
	Data not generated by user tracking	Device and software (e.g., operating system (OS), web browser)	The user is browsing the internet with a Samsung tablet, Android OS, and Firefox browser.
		Location of user	The user is in Paris, France.
		Date and time	The time of the user's visit is 2:00 pm on a Monday.
Publisher data (Non-User data)	**Ad slot-specific data**	Position of ad slot	The ad slot is on the top of the website.
		Ad format	The ad slot has a size of 728 × 90 pixels.
	Context-specific data	Content of publisher	The ad slot is on a website offering financial news written by professional journalists, such as www.financialtimes.com.
		Quality of publisher	The ad slot is on a website offering high-quality edited content, such as www.theguardian.com.
		Thematic-focus of publisher	The ad slot is on a finance and business news website like uk.finance.yahoo.com.
		Size of publisher	The ad slot is on the website of a small publisher, such as www.paris-normandie.fr.

Figure 1: Overview of data available with a typical bid request. Source: Laub et al. (2023)

2.1.3 Pricing models

Real-time bidding auctions happen based on the suggested costs per impression, usually in the format of CPM, or cost per thousand impressions, for easier calculations. However, as Tiet (2020) notices, there are several key pricing models employed by DSPs for display ads: cost per thousand impressions (CPM), cost per click (CPC) and cost per action (CPA), where action can be e.g. viewing the ad for some time, interacting with it or having some other signal of engagement with it. CPC and CPA were introduced in response to advertisers' demand for performance-oriented advertising (Tiet, 2020). It is especially these kinds of pricing models that the methods

presented in the current work are most applicable to, since they directly optimize the number of engagements or events per impression acquired.

As the industry moves towards attention-based metrics (Interactive Advertising Bureau and Media Rating Council, 2025), optimizing for events gains more relevance, since it could now be applied broadly to any advertiser that considers attention a good proxy for their campaign goals.

2.1.4 Bidding strategy optimization

To deliver ads in a real-time bidding setup, the fundamental task is to develop a bidding strategy, i.e. logic by which to bid on impressions. Ou et al. (2023) calls this task a "constrained optimization problem in an interactive and stochastic environment with big data as support". Indeed, an advertising engine can receive hundreds of millions of ad requests per day, which both enables a myriad of data science and big data techniques and puts strong pressure on inference efficiency. At the same time, there are diverse business objectives: some seek direct engagement from the user, such as click on the ad or conversion, while others seek long-term brand awareness. Moreover, there is a set of constraints such as total campaign budgets; budget pacing, which aims to spread the spending across the campaign timeline according to some pattern, such as even or front-loaded pacing; total return-on-investment (Ou et al., 2023).

Generally, the bid price depends on how well the ad matches the current user, and how competitive the auction is. In first-price auctions, the optimal bidding strategy is to bid slightly higher than the second-highest bid, but lower than the true value of the impression. Bid shading, i.e., bidding less than the impression is worth, is thus preferable but hard to carry out in practice because of the opaqueness of RTB auctions, where only the winner is notified of the winning price, which makes it difficult to collect data about the market. Ou et al. (2023) explains that it remains a challenging task to derive an optimal bidding strategy for value maximizers in first-price auctions. Several methods are proposed and could be used, such as a Proportional-Integral-Derivative (PID) controller that utilizes the real-time feedback to adjust bidding, as well as simpler rule-based approaches. For CPC-priced campaigns, the bid price could be, for example, computed as follows:

$$b = \frac{1}{\lambda} * CPC * p(click),$$

where λ is a coefficient regulating the desired delivery pacing of the campaign, lowering bids when the spending outpaces the desired behavior.

Unless the goal of the campaign is to simply deliver a maximum amount of impressions with minimum budget, any bid price calculation requires some estimation of the impression's true value to the bidder. The focus of this thesis is predicting the probability of a click on a given impression, which is an important signal of the impression's worth.

2.2 Tracking-enabled programmatic advertising

2.2.1 Advent of behavioral targeting

In the 2010s, the way people use the internet subtly but fundamentally changed. Before, the primary behavior on the web was search: users, usually via a desktop computer, used the web with a clear intent to find or purchase something. In contrast with this, the user experience on smartphones and tablets is focused on discovery. People use social media to catch up with their friends and follow sources they like, without a concrete purpose in mind, but rather just to browse content that is relevant to them. With the advent of infinite scroll feeds and short-form content, it became normal to consume content on the internet for hours on end, primarily on mobile but also on desktop.

On traditional TV, commercial breaks are standing between the viewer and the core media they are watching, which gives them no other option but to watch the advertisement. However, on the modern web, the core content is close and accessible to the consumer. As Busch (2015) points out, now web users "vehemently defend themselves against poor, irrelevant advertising" by scrolling past it or using ad blocker software. The publisher and advertiser need to ensure individual relevance in the exact moment the ad is seen, in order to grab enough attention for the person to read and consider it.

Programmatic advertising was one of the main ways in which the online advertising industry adapted to this environment. Armed with information about the person viewing the ad, adtech operators can collectively make sure the ad being shown is the most fitting to the user's intent. For example, the retargeting technique is widely used, where users who visited an e-commerce website but did not buy any product are targeted with ads of that website, encouraging them to return and finish the purchase. Breakthroughs in big data, machine learning and AI technologies enable more complex inferences about user's demographic characteristics, intent and susceptibility to specific advertisement topics or approaches, powered by big amounts of granular information collected by various brokers around the ecosystem, covered in more detail in Section 2.2.2. Generally, the promise of these technologies is that the more data is collected, the more accurate prediction is obtained and therefore better ad campaign performance is achieved.

Tracking and programmatic technologies are meant to make sure that the ads are showing something the users might actually want. This provides relevant suggestions to users and informs them about new products. Theoretically, fewer resources and less money are wasted on showing users irrelevant ads, and user experience of using the internet improves as a result. Most importantly in our context, on an individual level, tracking improves the revenue and metrics of both advertiser and publisher. Section 2.2.3 expands on the industry's case for behavioral targeting.

However, the practice of behaviorally targeted advertisements came under significant scrutiny in recent years. Section 2.2.4 explains how it violates consumer privacy, and harms the ecosystem of open web. This scrutiny prompted the EU, the UK, some US states and others to introduce legislation regulating and limiting these practices

(Section 2.2.5). In this environment, it is important to develop technology to show relevant, effective ads without infringing on privacy or harming the stakeholders. Consequently, contextual targeting emerged as a key tool to match ads and webpages in the absence of behavioral data, presented in detail in Section 2.2.6.

2.2.2 What is tracking

The main vehicle of cross-site tracking is so-called third party cookies. Cookies are small data files stored in the browser that help websites, for example, to keep users signed in. Mozilla documentation (MDN contributors, 2025) highlights three main purposes that cookies are used for: session management, personalization (such as language and UI theme), and tracking.

Cookies offer convenient persistent user identifiers, which enables adtech companies to collect and analyze individual data such as purchase, browsing and location history. Even in absence of cookies, user identification can also be achieved with device fingerprinting, based on signals such as IP, MAC-address of the device, and other metadata.

Other tracking technologies include pixels – code injected as invisible, transparent `1 x 1` images that are e.g., embedded in an email, which collects information about whether the individual user opened the email and when, as well as whether they clicked any links in it. Tracking pixels also offer additional data on the user's behavior on the page, including how much time they spent on the page and where they clicked, etc., which can be used to infer data about the magnitude of interest and engagement with the content achieved, including how much attention the user paid to the ad, even if they did not click it. Fouad et al. (2018) found that this technology was used by 83% of web domains they surveyed.

Persistent identifiers described above enable advanced profiling of users. Demographics, interests and political beliefs can be inferred, which are then made into granular ad segments available for advertisers to use for targeting. Figure 2 shows the process of behavioral targeting, from behavioral signal to ad impression.

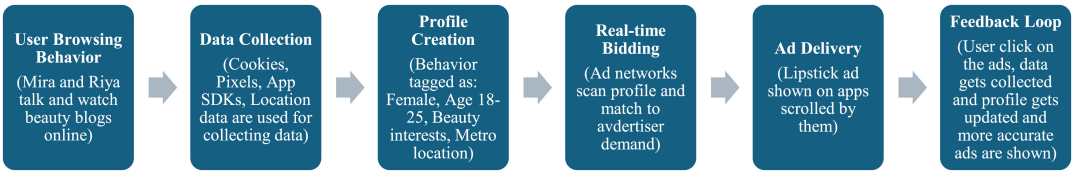

Figure 2: Illustration of behavioral targeting flow. Source: Wadhwa and Bachwani (2025)

Even more granular, hyper-personalized targeting optimizations can be achieved by directly estimating the probability of conversion or engagement from the data collected. Machine Learning approaches such as collaborative filtering could be used, which assumes that users with similar characteristics behave similarly. Some of these approaches are covered in Sections 2.3 and 3.

2.2.3 Value of tracking

So how much value does tracking actually bring to advertisers and publishers? Why do most publishers enable third-party cookie technologies on their websites?

Empirical evidence to answer these questions started appearing as regulators in several countries enforced a possibility for every internet user to opt out of usage of third party cookies, thus making tracking some users much more difficult. In particular, several groups of researchers measured the publisher revenue lift from tracking technologies such as cookies. Conventional thinking is that when targeting is better, the advertisers are able to spend more per impression, driving up the ad prices and thus publisher revenue. Most found a statistically significant difference in prices of ads with targeting available and without; however, the magnitudes are very different.

Laub et al. (2023) carried out two separate studies using browsing traffic in 2016 and 2023, across mobile and desktop, as well as EU and US. Both studies found that "ensuring user privacy online has substantial costs for online publishers". In particular, the researchers found that average ad price decreases 18% (Study 1) and 23% (Study 2) when user tracking is unavailable, although there was variation between publisher groups.

Gu et al. (2026) cover an experiment jointly overseen by Google and UK Competition and Markets Authority (CMA). The analysis of results found that unavailability of third party cookies results in publisher revenue dropping by 29.1%. The paper also recorded that Google Privacy Sandbox, a more privacy-friendly tool to replace cookies, recovers only 4.2% of that lost revenue, though this number might be affected by generally low adoption of this tool.

However, some highly cited papers arrived at different numbers. Marotta et al. (2019) carried out an analysis of data from multiple websites with correction for potential confounding features, which yielded a finding that cookies only bring 4% of additional revenue.

P. Wang et al. (2024) took advantage of the GDPR compliance rollout across European publishers to record the impact to their business metrics. Analysis of billions of impressions showed a modest drop of 5.7% in publisher revenue, explained in part by some advertisers not bidding at all when a tracking identifier is not available.

For an individual publisher, these studies offer a clear financial incentive to enable tracking pixels and collection of third-party cookies, especially considering the plug-and-play ease of their implementation. However, when these technologies are enabled by majority of web publishers, they might hurt the collective revenues more than the measured lift they provide, on top of obvious negative privacy impacts, as illustrated further.

2.2.4 Negative externalities of tracking

A difference of almost an order of magnitude illustrates diverging views on behavioral targeting in the literature. The adtech industry maintains that behavioral targeting is a win-win situation for advertisers, publishers and consumers alike. Theoretically, more data results in smarter predictions, which gives benefits such as lower search

costs and better matching between advertisers and consumers. However, scholars have found numerous inherent negative economic effects on the advertising ecosystem and stakeholders associated with tracking. Big adtech companies end up in a position where they are able to carry out anticompetitive practices, diminishing the revenue flow that powers the free internet. Moradi et al. (2025) formulates these two views of online advertising, illustrated by Figure 3.

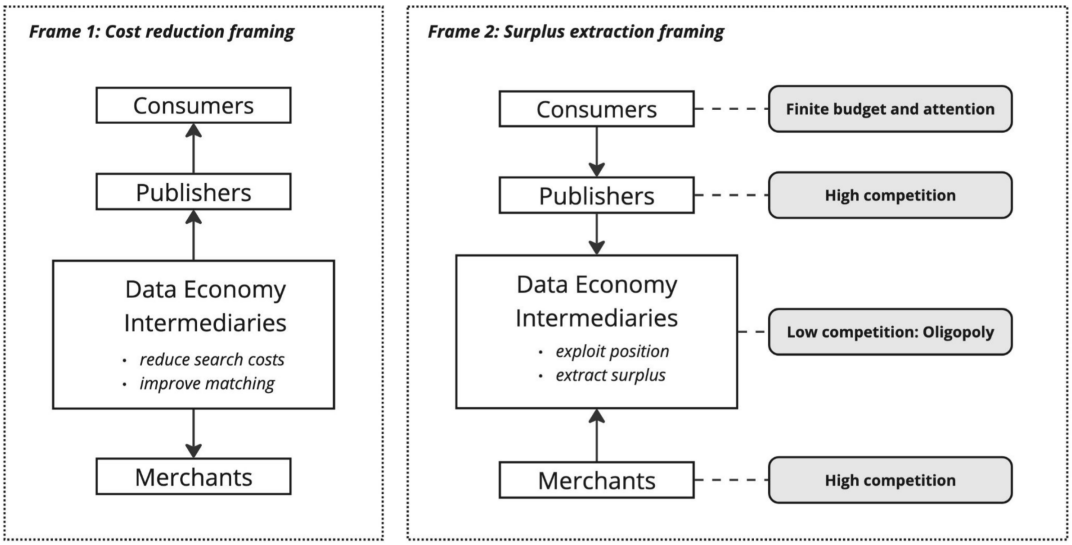

Figure 3: Two framings of the role of data intermediaries in online advertising ecosystem. Arrows show how value is either created (frame 1) or extracted (frame 2). Source: Moradi et al. (2025)

Indeed, consider the nuanced impact of the tracking availability for advertisers. Powered with behavioral targeting, merchants can reach the "right" consumers, paying for impressions that will most likely result in conversion. Without tracking, however, a producer of skiing equipment may have historically targeted winter sports sections of online newspapers, where it competed with other businesses in related industries for impressions and consumer attention. Since tracking enables targeting using people's inferred preferences, interests and demographics, the merchant is now competing against an unbounded collection of advertisers of products that might be relevant to a particular visitor of the webpage, be that luxury travel, football or professional education. A similar dual dynamic can be traced on the publisher side. On one hand, the content is more easily accessible for bidding by various merchants, which makes more impressions happen, which convert to publisher revenues. On the other hand, behavioral targeting erodes the value of publishing specialized content, which is a tool that lets publishers achieve differentiation from competitors. In other words, the power to match merchants and consumers shifts completely from web content owners to data intermediaries, such as Google.

In context of limited consumer attention and budgets to spend on advertised products, tracking intensifies competition between publishers and between advertisers. Consequently, it reduces the bargaining power of both sides, which puts data

intermediaries in a uniquely strong position. Adding the fact that effective behavioral targeting demands vast amounts of data collected across platforms, which restricts this technology to be offered by big players only, tracking enables an oligopoly, as demonstrated by anti-trust lawsuits against Google. In some prominent cases, the intermediaries were able to charge as much as 70% of the advertisers' budgets, leaving less than a third to web page owners (Davies, 2017).

Advertising powers the Open Web, or the freely accessible portion of the internet, such as newspapers and blogs without a paywall. Revenues from advertising traditionally support the platforms offering democratic access to information, such as independent local newspapers that are a primary source of news to many, offering a sustainable alternative to subscriptions and content restricted for subscribers, which reduces the reach and impact of the articles and is only viable for big publishers. In the United States, almost 40% of local news outlets have shut down since 2005 (Metzger, 2025), consistent with tracking technology having a negative impact on the financial situation of publishers.

Negative economic consequences of tracking motivate development and usage of the machine learning approaches such as the one introduced in this thesis, for publishers and advertisers. To the author's knowledge, contextual targeting has not received significant attention in data science research. If that was to change, the gap between advertising outcomes highlighted in the previous section could be narrowed or closed, which would motivate individual publishers to opt out of third-party tracking altogether and reverse its destructive effects on the ecosystem. However, beyond strictly economic effects, tracking also has important ramifications for stakeholders such as individual internet users.

Importantly, tracking and vast data collection harms privacy of individuals. Any one of hundreds of participants in real-time bidding auctions receives full data on impressions, including a personal identifier, even if they do not actually send a bid. This way, it is impossible for an individual to know where their precise location data or browsing history might be collected and stored. Moreover, this vast data collection happens without an explicit opt-in consent (except cookie banners). Potential harms associated with violation of privacy are very real; one way to map them is presented in Table 1.

Behavioral targeting also has some inherent effects on individuals. Summers et al. (2016) studied how when consumers realize that the ads are being targeted based on an inference made about their identity, this inference becomes the implied social label, influencing the way individuals perceive themselves.

2.2.5 Data protection laws

Because of the harms illustrated in the previous section, countries around the world introduced legislation to limit data collection, retention and sharing practices.

Perhaps the most consequential data protection law is the General Data Protection Regulation (GDPR; Wolford, 2018), which is a law protecting people in the European Union. This law enforces several data protection principles, such as purpose limitation; data minimization; storage limitation and confidentiality. Data processors must notify

Table 1: A typology of privacy harms as per Citron and Solove (2022).

Type of privacy harm	Subtype	Example
Physical harms		Data brokers giving victim's home address to criminals; doxing
Economic harms		Identity theft as result of a data breach, leading to fraud
Reputational harms		Incorrect data records leading to failing credit report
Psychological harms	Emotional distress	Fear from personal information leak
	Disturbance	Telemarketing calls
Autonomy harms	Coercion	Higher prices for those exercising privacy rights
	Manipulation	Cambridge Analytica influencing elections
	Failure to inform	Background check carried out without informing the subject
	Thwarted expectations	Selling of personal data to third parties
	Lack of control	Retention of personal data beyond maximum period allowed/expected
	Chilling effects	People not engaging in dating apps anymore because of one of them admitting to selling private information
Discrimination harms		Hiring service using intimate information to score candidates
Relationship harms		Disclosure of confidential communications

the data subjects of a data breach within 72 hours or face steep penalties. Recent regulation proposals attempt to reduce cookie fatigue and clarify usage of personal data for AI model training (European Commission, 2025).

In the United States, California has passed the California Consumer Privacy Act (CCPA), which is similarly consequential as many global technology companies are affected by it. The law secures the privacy rights such as the right to know how the personal information is collected; the right to delete personal information collected and the right to opt-out of sale or sharing of the personal information (California Department of Justice, Office of the Attorney General, 2024).

2.2.6 Contextual targeting

The laws introduced in Section 2.2.5 and associated compliance costs and risks create additional pressure for participants of the online advertising ecosystem to consider alternative ways of targeting. Contextual targeting - optimizing advertisement effectiveness by placing it in favorable media contexts (K. Zhang and Katona, 2012) - has emerged as a privacy-friendly option.

In their analysis of the impact of GDPR compliance on publisher revenues, P. Wang et al. (2024) notes that ads placed on web pages with relevant topics contributed to 40-45% recovery of the conversion rate lost with unavailability of user personal data. In other words, relevant context softens the blow when tracking is unavailable.

Häglund and Björklund (2024) describes how a "shift in data supply" motivates increasing use of contextual advertising. Personal data is becoming more difficult to handle because of developing regulations and emerging privacy features, while contextual data is getting easier to process due to advances in natural language processing (NLP) and artificial intelligence (AI).

The researchers focus on three context factors influencing advertisement effectiveness. *Applicability* denotes the fundamental way in which the content around the ad placement is relevant to the ad content. There are nuances in how this applicability impacts the perception of the ad. In the example used, an ad for a large car was perceived positively in the context of safety, but it was perceived negatively in the context of fuel economy. *Affective tone* is the way sentiment of the article and ad influences its perception. Finally, *content involvement* is a context factor motivated by research showing that engaging content results in better ad recall and conversion rates. Häglund and Björklund (2024) then show how AI can be applied to detect all three of these factors.

2.3 Machine Learning in Programmatic Advertising

2.3.1 Recommender systems

The task of deciding which ad to show to which user can be reduced to the recommendation task: based on a trove of data about past individual behavior of a big number of users, rank the options in order of potential preference of a specific user. Recommender systems – machine learning systems designed for this task – are widely used in e-commerce, entertainment, tourism and other industries, as well as in advertising technology. A timeline of development of the recommender systems is shown in Figure 4.

The foundational idea employed in most recommender systems is Collaborative Filtering (CF). CF is a technique used to predict the preference of a user based on known preferences of similar users. In its basic form, it works with a user-item matrix, where past user interactions fill the corresponding cells. However, such a matrix is usually very big and sparse, especially when there are thousands of items and each user has interacted with just a few of them on average. This makes it hard to use for meaningful predictions directly and prohibitively expensive in terms of computation

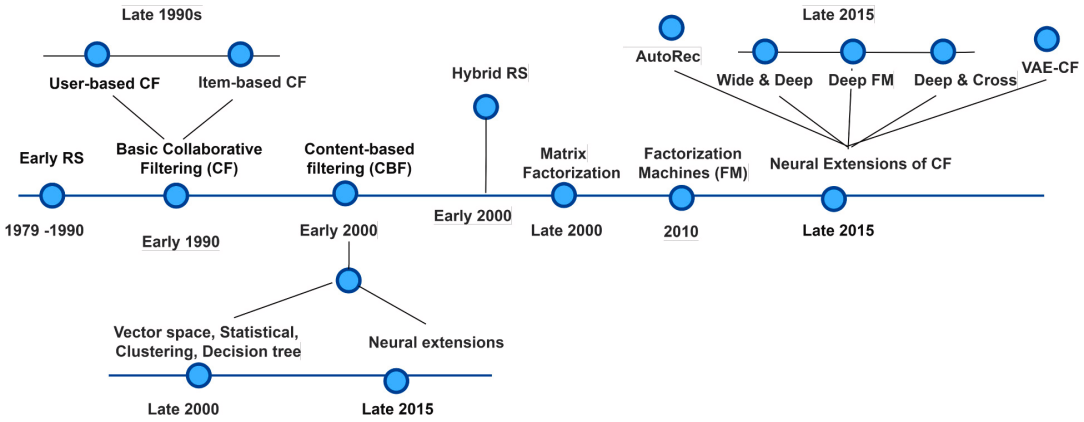

Figure 4: Progress of research on key recommender systems. Source: Raza et al. (2026)

and memory. Thus, the recommender systems research focused on ways to reduce this huge matrix to smaller ones, while retaining and condensing useful information. One widely used approach is Factorization Machines (FM), introduced in Section 3.4. FM, as well as other modern approaches, is also able to incorporate information about features of items, users and interactions, not just about the interactions themselves. Thus, the latest research has been focusing on how to model feature interactions in a way that makes compute complexity manageable while maintaining important information. Approaches such as Wide & Deep (Cheng et al., 2016) try to preserve both low-level, pairwise feature interactions and high-level patterns that pure FMs are not able to model efficiently.

2.3.2 Language Models

In the last decade, data science research made serious advancements in encoding understanding of natural language texts. In their foundational paper, Vaswani et al. (2017) introduce the Transformer architecture, illustrated in Figure 5.

The model was originally designed for translation tasks and consists of an encoder and decoder that use stacked self-attention and fully connected layers. "Attention Is All You Need" (Vaswani et al., 2017) suggested using self-attention instead of convolutional or recurrent layers for the first time. The researchers point out advantages such as the ability to parallelize computations, since matrix multiplications are used instead of sequential recurrent operations. Moreover, this architecture was found to be better at learning long-range dependencies in text due to shorter paths between any combination of tokens in input and output sequences.

Devlin et al. (2019) took this idea further by introducing the Bidirectional Encoder Representations from Transformers (BERT) model. The main contributions of this paper are showing how to train and use a *bidirectional* transformer, and making Transformer architecture conveniently usable for a wide range of downstream tasks besides translation. In the original Transformer paper, each token attends only to previous tokens in self-attention layers due to the nature of training process, which

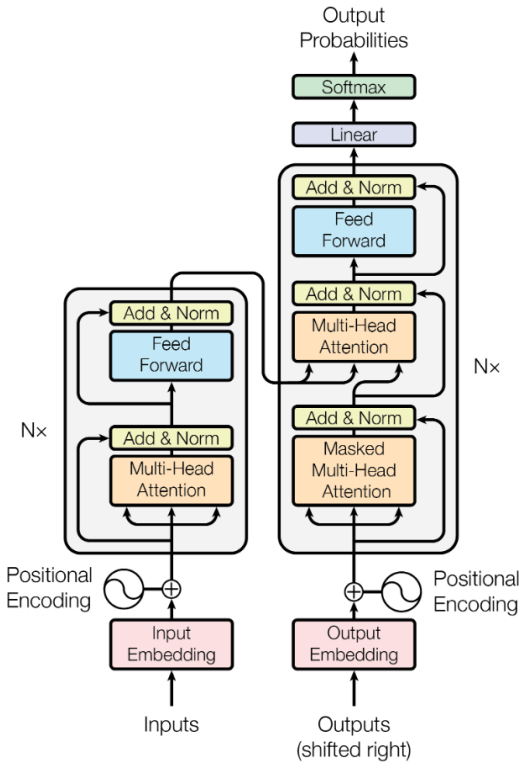

Figure 5: Transformer architecture. Source: Vaswani et al. (2017)

constrains the representative power of the model. BERT tackles this constraint using a masked language model (MLM) pre-training objective, where random tokens from the input are masked, and the model's goal is to predict the masked tokens based on the context around them. In addition to MLM, the next sentence prediction (NSP) task was used to train the transformer, so all inputs contained also a special `[SEP]` separator token and segment embeddings. The resulting pair of "sentences" were either occurring one after the other in the training text or randomly sampled, and the model was trained to distinguish the two using its special `[CLS]` hidden state vector. BERT input format is shown in Figure 6. Note that sentence in this context means a portion of text, not necessarily a grammatical sentence.

$$\texttt{[CLS]} \;\; A_1 \ldots A_N \;\; \texttt{[SEP]} \;\; B_1 \ldots B_M \;\; \texttt{[SEP]}$$

Figure 6: BERT input format for a sentence pair: both sentences packed into one sequence, with `[CLS]` at the front and `[SEP]` marking the boundaries. For single-sentence tasks such as text classification the second sentence is simply kept empty.

This design allowed researchers to train a model with a deep understanding of natural language, useful for various downstream tasks such as classification, question-answering, entity recognition and others. Importantly, high performance on these tasks can be achieved without designing a task-specific architecture, but simply by

fine-tuning some layers of BERT model and adding an output layer on top. `[CLS]` especially proved useful for e.g. binary classification tasks. Finally, the encoder-only architecture of BERT is generally more computationally efficient for representation tasks such as CTR classification.

Later works, notably Robustly optimized BERT approach (RoBERTa), suggested by Y. Liu et al. (2019), improved the training hyperparameters and dropped the NSP task at pre-training as it was not found to add performance. The model has, however, retained an ability to fine-tune on pairs of sentences, to the advantage of the current thesis. XLM-R, a cross-lingual language model based on the RoBERTa architecture (Conneau et al., 2020), is a version of RoBERTa trained on two terabytes of data in 100 languages, bringing the deep natural language understanding to use cases outside of English-speaking countries.

2.3.3 CTR prediction

Click-through rate (CTR), or simply click prediction, is one of the central tasks in advertising technology. It measures the probability of a display ad being clicked when shown and it is used to estimate the value of an impression and consequently the bid price. Common CTR values for desktop display advertising are below 0.1% (W. Zhang et al., 2014).

The simplest way to predict CTR for bidding purposes is a rules-based system which, for example, extrapolates past statistics for a specific ad format. However, the high-volume nature of real-time bidding and straightforward formulation as a binary classification task offer opportunities to use machine learning approaches to achieve better performance. Strict inference time constraint and high class imbalance (about 1 click in 400 impressions) make it a challenging problem.

Historically, logistic regression was once a popular tool for the task, for its simplicity and interpretation, as well as low computation cost. After that, several types of recommender systems have been used as more data and compute were available: namely, factorization models, deep learning models and hybrids, as shown in Figure 4.

Since the problem statement of the current thesis does not allow the use of any data traceable to individual users, we are unable to utilize the collaborative filtering assumption and are especially interested in methods that put an emphasis on contextual features, including data about the ad and the article. Intuitively, the ad should address the same interest or intent which brought the user to the page where the ad is shown.

In this regard, it is useful to consider models designed for a related problem: CTR prediction in search advertising. Search engines use CTR predictions to provide relevant sponsored placements in the search results. (Notably, the context considered here is search engines before they started to use AI for features such as generated summaries and direct question answering.) Typically, it is important that paid suggestions are relevant to the search query, which is similar to our goal of making a display ad relevant to the web page it is shown on. The BERT-CTR model developed in this thesis is based on one such recommender designed for sponsored search (D. Wang et al., 2023).

2.3.4 Key evaluation metrics

CTR prediction is essentially a binary classification task with two classes: click and no click. Because of very high class imbalance ($\approx$ 1:400 ratio of clicks to non-clicks), accuracy and related performance metrics are not usable, since a naive model always predicting $p(click) = 0$ would yield an almost perfect score. Instead, CTR prediction literature commonly uses two more robust metrics: log loss and AUC (Alotaibi and Alotaibi, 2025).

Log loss, or binary cross-entropy loss, is defined as follows:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right]$$

where y_i are ground truth labels from the training dataset, and $\hat{y}_i$ are the model's predictions. Log loss measures the difference between the distribution of predicted scores and real labels. Smaller values mean better performance. Importantly, log loss penalizes confident but wrong predictions, which is important in CTR prediction since it alleviates the chance of overpaying for some impressions.

AUC measures the probability that a randomly selected positive item ranks higher than a randomly selected negative item, also formulated as the area under the ROC curve (Bai et al., 2025). A random classifier yields AUC of 0.5, while 1 is the perfect score. AUC is a metric insensitive to class imbalance.

Relative information gain (RIG) shows the difference in log loss compared to a baseline, as follows:

$$\text{RIG} = 1 - \text{NormalizedEntropy} = 1 - \frac{\mathcal{L}_{\text{model}}}{\mathcal{L}_{\text{baseline}}}$$

In this thesis, as in D. Wang et al. (2022), RIG is used for presentation convenience, to show the performance improvement of models. Often the baseline in RIG is a naive fixed-CTR predictor, but, since a slightly less naive and more meaningful status quo predictor is available, this work uses it as the baseline instead. One advantage over the absolute log loss difference is that it is independent of the base rate of click probability.

3 Methods

To answer the research questions, I compare three models of increasing complexity against a production rule-based baseline (RQ1, RQ2), and separately evaluate the developed model against practical constraints (RQ3).

3.1 Data

In the absence of personal data, we need to utilize all contextual information available. Contextual signals for prediction of a click include 1) temporal features, such as hour of day and day of week (perhaps, some ad is more effective late at night or during the workday); 2) client context, such as whether the ad is shown on a mobile or desktop device, and whether it is shown on an Apple device or an often cheaper Android device; 3) ad context, such as type of creative (e.g., video or text) and textual description of the product being advertised; and 4) webpage context, such as title and excerpt of the page, as well as position of the potential ad relative to other content.

The data for this thesis has been collected using internal infrastructure of an advertising technology company, a demand-side platform called C Wire. C Wire is a participant of online display advertising ecosystem in Switzerland and neighboring countries. The company focuses on delivering advertisements online where third-party cookies are unavailable.

Models were trained and evaluated on historical data collected in the period from February 1 to March 31, 2026. Each datapoint is a single event of an advertisement rendered and shown to a user, also called an impression. The dataset contains 15.2M impressions. The full list of columns suitable for use as features is shown in Table 4. Click, specifically the event of the user clicking on the ad and thus opening the advertiser's link, was taken as the target for prediction and the label for supervised learning. Other relevant events could, however, also be used, such as time on screen metrics or non-click interactions with interactive ads, which could signal attention spent on the ad.

It is worth noting that profiling is one approach sometimes used in the advertising industry in an environment where conventional personal tracking is not available. This technique uses granular information about the client and connection that still makes it to the adtech provider, such as IP address, browser version and viewport dimensions, to deanonymize the user and connect impressions of the same user, possibly with other known personal information. Resulting deanonymized data could be used with mainstream personal-data-based click prediction approaches. However, this approach violates privacy of internet users and often betrays their explicit choice to opt out of personalized advertising, and so is not a part of this thesis.

All of the methods described in the following sections used 80% of this dataset for training and the remaining 20% for validation. The train-test split was not carried out randomly, but rather chronologically: the first 80% of impressions were placed in the training set, while later impressions were held out as the validation set. This makes evaluation closer to a real-world scenario and avoids leakage of future data into training.

3.1.1 Dataset statistics

Table 2: Basic dataset statistics.

	Train	Test	Total
Impressions	12,169,368	3,042,345	15,211,713
Clicks	29,574	7,769	37,343
Click-through rate	0.2430 %	0.2554 %	0.2455 %
Start	2026-02-01	2026-03-22	2026-02-01
End	2026-03-22	2026-03-31	2026-03-31

The dataset is characterised by heavy class imbalance: a click happens for every ≈400 impressions. Negative downsampling could be used for training efficiency (taking fewer examples of the negative class and correcting the prediction accordingly post-hoc). It was not used in the current thesis, since calibration of the models was crucial for the requirements. Moreover, training costs were manageable, and the loss function and metrics are agnostic to class imbalance. Future work could experiment with using downsampling for training the models in this thesis.

3.1.2 Textual data

The articles are scraped using Mozilla's `readability` library. In particular, `title` and `excerpt` fields are retrieved. This library is an industry standard, in part because its speed is suitable for extremely high data volumes of adtech. However, it often comes at the cost of data quality. Designing a scraper was out of scope for the current thesis, but there might be significant potential improvements to be found from a better dataset of article texts.

Ad texts were not readily available and were scraped as part of the current work. The commercial LLM Claude Sonnet 5 was used to scrape and summarize the call-to-action (CTA) URL of the advertisements. The prompt used can be seen below:

System Prompt for summarizing the landing page of advertised product

```
Based on the web page content below, write a single short
sentence naming the product or service being advertised and
the company offering it.  If the page does not clearly advertise
a specific product or service (e.g.  it is a generic login page,
a cookie/consent interstitial, an error page, or the content is
too sparse), respond with exactly:  unclear

URL: {final_url}
Title:  {title}
Body excerpt:  {body_excerpt, first 2000 chars}
```

In the textual part, the ad text (scraped description of the advertised product) is encoded with the title and excerpt of the webpage as a BERT sentence pair (in original

paper, search query and ad text; see Figure 6). This way, during fine-tuning, the transformer model learns which combinations of ads and articles are most effective. Before training and inference, textual pairs are truncated at 128 tokens (≈0.5% records affected), longest-first. Distributions of token lengths in the data are noted in Table 3.

Table 3: Token length distributions

Field	Mean	P50	P90	P95	P99
ad_text	21	20	33	38	39
title	12	12	18	22	29
excerpt	33	31	46	55	68
title+excerpt	45	43	62	74	89
combined pair	70	68	92	102	121

Table 4: Features used in training. LR/FFM use the tabular features only; BERT-CTR additionally consumes the text fields.

Feature	Type	Card.	Models	Description
Temporal				
hour_sin	dense	cont.	all	$sin(hour \cdot 2\pi/24)$
hour_cos	dense	cont.	all	$cos(hour \cdot 2\pi/24)$
day_of_week	categorical	7	all	day name
Client context				
Endpoint_OS	categorical	7	all	Client OS (iOS 46.7%)
Endpoint_Country	categorical	101	all	Mainly Switzerland
Ad context				
Creative_Type	categorical	6	all	text / story / video
Creative_Variant	categorical	16	all	creative layout
ad_text	tokenized	356	BERT	LLM summary of the product
Webpage context				
publisher	categorical	23	all	Publisher
pl_position	categorical	12	all	Ad position on the page
title	tokenized	46,616	BERT	Article title
excerpt	tokenized	46,996	BERT	Article excerpt
Historical prior				
ctr_baseline	dense / binned	64 bins	all	Hierarchical historical CTR (3.2)
Label				
HasOpenedTarget	binary	2	all	User opened the advertised target.

Note: hour was encoded with sin/cos to preserve its cyclical nature and the close relationship between 23:00 and 00:00.

FFM uses binned `ctr_baseline` because the implementation used (libffm) does not normalize dense features, which would give a small weight to a feature that generally has such low values.

3.2 Baseline: rule-based algorithm

As a baseline, I took a historic average click rate based on data available. In particular, the CTR was computed over a 60-day lookback window, with three hierarchy levels:

- CreativeType level, most broad
- (CreativeType, PlacementId)
- (CreativeType, PlacementId, LineItemId), most specific

For each impression, the baseline CTR prediction was computed over the most specific level that had over 1000 impressions accumulated. For example, for a campaign that just started, the CTR could be computed as a count of all clicks recorded on the same placement by creatives of the same type in the last 60 days, divided by the count of all impressions recorded in those conditions. This way, we ensure that the empirical CTR is as close to the expected click probability as possible with a reliable leverage.

3.3 Logistic Regression

The logistic regression model is one tool which has been used for CTR prediction (Richardson et al., 2007). As a shallow model, logistic regression offers interpretability and low complexity compared to other models, at the cost of limited predictive power. Richardson et al. (2007) applied this approach to search advertising and used carefully engineered features, such as bid term statistics and 81 manual ad quality features. This highlights how performance of logistic regression is sensitive to feature engineering, because the model assumes a linear relationship between input features and the logits of the outcome, and it is unable to model feature interactions.

Logistic regression uses the following formula for its predictions:

$$\hat{y} = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}}$$

Here, $\hat{y}$ denotes the predicted variable (in our case, probability of a click), $\mathbf{w}$ is the learned weights vector and $\mathbf{x}$ is a vector of features in numeric form. In this thesis, logistic regression was trained with a set of dense features as well as one-hot-encoded categorical features described in Table 4. Textual features are not directly applicable to this shallow model.

The model was trained using the classic limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method (D. C. Liu and Nocedal, 1989). Cross-entropy loss (also called log-loss) was used, which is defined as follows:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{1}$$

3.4 Field-aware Factorization Machine (FFM)

Factorization machines (Rendle, 2010) are a class of models designed for huge and sparse data, with the additional benefit of modeling pairwise feature interactions. These qualities proved to be very suitable for the ad targeting context, when a lot of data is mined for each user, and a new inference often needs to be processed in milliseconds.

The core idea of the algorithm is to avoid having a matrix of n^2 weights for feature interactions by maintaining n latent feature vectors, the dot product of which would be the weight for the corresponding feature pair. This idea is somewhat similar to matrix factorization, which is why it is called a factorization machine. In more formal terms, the model equation is as follows:

$$\phi_{\mathrm{FM}}(\mathbf{x}) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle \mathbf{v}_i, \mathbf{v}_j \rangle \, x_i x_j \tag{2}$$

where the model parameters that have to be estimated are:

$$w_0 \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^n, \quad \mathbf{V} \in \mathbb{R}^{n \times k}$$

And $\langle \cdot, \cdot \rangle$ is the dot product of two vectors of size k:

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}$$

Rendle (2010), which introduced the model, also suggests a way to construct a "d-way Factorization Machine" that would model d-way feature interactions. However, as Guo et al. (2017) points out, the computational complexity grows significantly, limiting the real-world usage to d=2.

When a lot of past decisions by users are known (for example, whether they went to a certain product webpage), millions of sparse features are created via one-hot encoding. Thus, significant research efforts were spent on trying to model more complex feature interactions in an attempt to increase the predictive power of factorization machines, while taking advantage of their light computational costs. In particular, DeepFM (Guo et al., 2017), DCNv2 (R. Wang et al., 2021) and Wide & Deep (Cheng et al., 2016) became some of the state-of-the-art recommender systems, combining the expressiveness of deep neural networks and the efficiency of factorization machines. These are widely used in advertising, when a lot of sparse data is available, which makes CTR prediction a recommendation task.

In a privacy-preserving context, however, large datasets of individual choices are not available. FM-like models are often designed for millions of features, while the data examined in this thesis only offers thousands. However, it is still worth it to evaluate the performance of this class of models in this environment.

The model chosen in this thesis is called Field-aware Factorization Machine, introduced by Juan et al. (2016). The approach took several top spots in public

contests by major adtech companies such as Criteo and Outbrain, and its code is readily available on Github.

In FFM terminology, a column of categorical data such as Endpoint_OS in our dataset is called a "field". Since the model cannot process values such as "iOS" and "Windows" directly, we first need to convert this column into numerical features. Usually this is done using one-hot encoding, in which each possible value of the column gets its own feature that takes a value of 1 when this value appears in this column and 0 for other rows. The intuition of FFM is that interactions between different OS-derived features are different in nature than interactions between the OS column and e.g. day of week.

While in FMs weight computations use a per-feature latent vector v, in FFMs there is a latent vector for each feature-field pair. In formal terms,

$$\phi_{\text{FFM}}(\mathbf{w}, \mathbf{x}) = \sum_{j_1=1}^{n} \sum_{j_2=j_1+1}^{n} (\mathbf{w}_{j_1,f_2} \cdot \mathbf{w}_{j_2,f_1}) x_{j_1} x_{j_2}$$

Taking n as the number of features and f as the number of fields, the total number of parameters rises to nfk from FM's nk, so the authors recommend taking a lower k value. In this thesis, a random hyperparameter search was carried out in order to find the best model configuration for the given dataset, details of which can be found in Table 5, with results demonstrated in Figure 7. All runs used early stopping as recommended by the authors.

Table 5: Configuration of FFM hyperparameter search.

Parameter	Distribution	Range
n of latent factors k	uniform discrete	2, 4, 8, 16
learning rate r	log-uniform	[0.01, 1.0]
L2 regularization parameter l	log-uniform	[1e-6, 0.1]
instance normalization flag no_norm	bernoulli	true, false

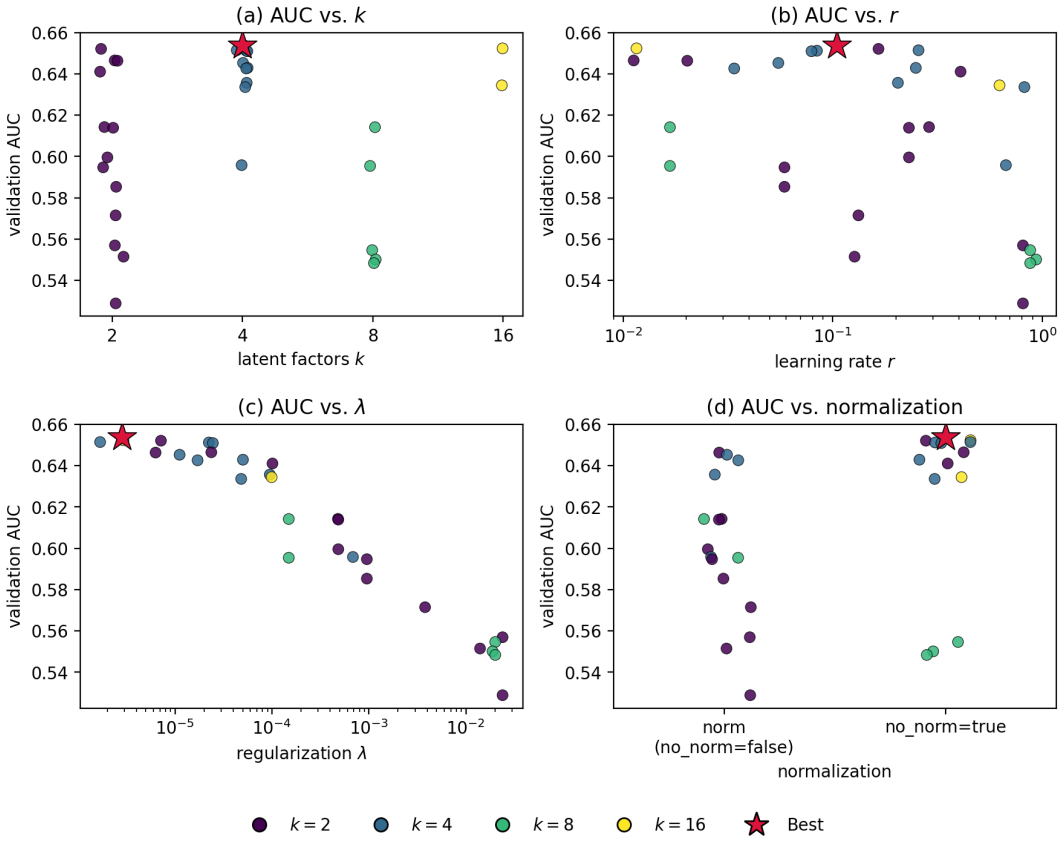

Figure 7: Results of FFM random hyperparameter search (n=30).

3.5 Late fusion BERT-CTR

Intuitively, an important signal of ad effectiveness is its relevance to the content around it. Indeed, the user is already interested in the topic of the web article they are reading, so it makes sense to show ads related to that interest. However, the methods covered above do not offer a productive way to capture the complex connections between the ad and page content and their impact on click probability. At the same time, the transformer architecture, introduced in the foundational paper by Vaswani et al. (2017), has recently demonstrated advanced understanding of natural language in a range of tasks.

Naturally, the numerical contextual signals still carry a significant signal, which should be incorporated into predictions alongside text. D. Wang et al. (2023) compares several ways to achieve this blend. One technique is to simply concatenate the numbers as textual tokens with actual text, and feed it to the pre-trained language model as input (X. Zhang et al., 2020). However, this approach was found to underperform compared to others, while having large requirements for inference, since computations over the numerical data now unnecessarily include the whole transformer, instead of optimized models like FFM.

D. Wang et al. (2022) notably use a pre-trained language model (LM) for CTR prediction in the sponsored search use case for Microsoft's search engine Bing. The

architecture combines a textual tower, with a fine-tuned BERT-family LM, and a non-textual tower with embeddings over non-textual features, with node-wise summation providing the fusion of features and resulting in a low-dimensional set of numeric features to be fed into a subsequent CTR prediction pipeline.

In the current thesis I use a variant of this architecture that functions as a standalone CTR prediction model, suggested by D. Wang et al. (2023) in their follow-up paper (referenced as "Shallow Interaction architecture" baseline), illustrated in Figure 8. Section 5 explains why the paper's flagship BERT4CTR model was not considered in the experiments of this thesis.

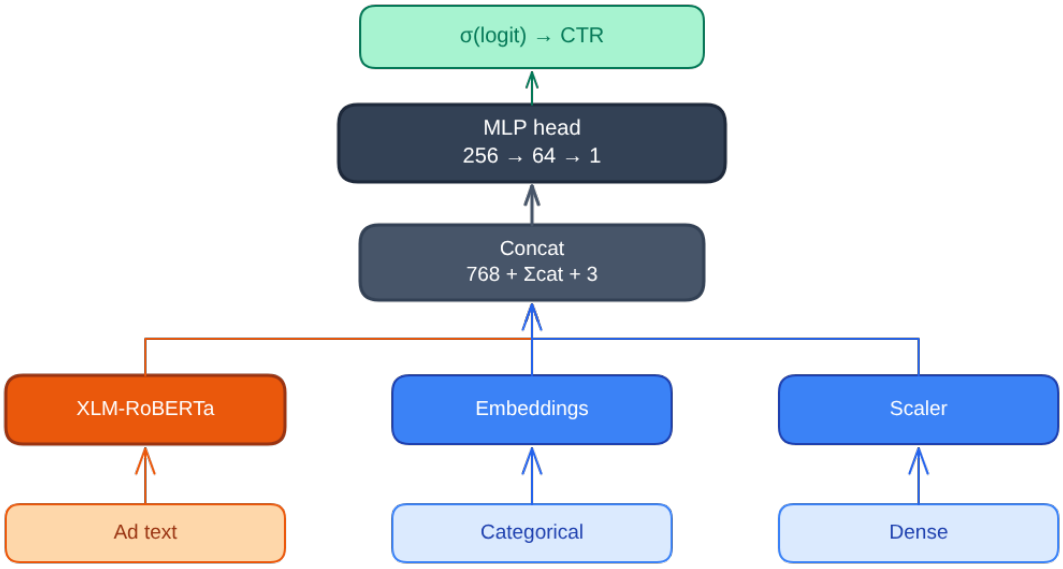

Figure 8: Late fusion architecture used.

The pre-trained transformer model published as `xlm-roberta-base` is used, a version of the XLM-R model introduced in Section 2.3.2, since the corpus is multilingual. Dense features are z-score standardized and concatenated as raw scalar floats into the fused vector, as opposed to being 101-dim one-hot-encoded in the original paper.

3.5.1 Training details

The training setup includes data processing, which aggregates the ad events data by impression, removes the impressions for which not all training data is available, performs data augmentation and cleaning. Training experiments are monitored and tracked using the open-source framework MLflow (mlflow.org). Various memory optimizations are performed in the training code to keep the requirements feasible despite large datasets. The main training job of BERT-CTR model has 150GB RAM requirement, while data preparation requires up to 300GB of available memory. The high-performance `polars` Python library and rigorous batching are used.

The training was carried out on Nvidia A100 and H100 GPUs using the Triton

high-performance cluster of Aalto University. Training architecture, optimization choices and hyperparameters are shown in Table 6.

Table 6: BERT-CTR model training hyperparameters.

Hyperparameter	**Value**
Architecture	
Backbone	xlm-roberta-base (hidden size 768)
Frozen backbone layers	embeddings + lowest 8 of 12 blocks (top 4 trainable)
Cat. embedding dim	$\min(50, \max(4, \text{round}(1.6\, C^{0.56})))$, where C - cardinality
MLP head hidden dims	256, 64 (ReLU)
Dropout	0.2
Output bias init	$\text{logit}(\text{base rate in training set})$
Max sequence length	128
Optimization	
Optimizer	AdamW
Loss	Binary Cross-Entropy with logits
Backbone learning rate	1×10^{-4}
Head learning rate	Learned in hyperparameter sweep per variant, see Figure 10
LR schedule	linear warmup $\rightarrow$ linear decay to 0
Warmup fraction	0.1 of total steps
Weight decay	0.01
Epochs	2
Effective batch size	2048
Gradient accumulation steps	4 (micro-batch 512)
Train/test split ratio	0.8 (temporal)

Note that the experiments found that the model had a high variation in validation loss and AUC from training initialization (seed). In particular, since no downsampling was done, high batch size means that there is a high probability that almost every batch will have at least one positive datapoint, resulting in a productive gradient update. Warmup of learning rate stabilizes early BERT fine-tuning and MLP training by gradual ramping up of the learning rate over the first 10% of steps. This way the model is less sensitive to random weight initialization. Finally, the bias of the MLP output layer is also initialized to the logit corresponding to the training set ratio of positive samples. This makes the training start with weights already roughly reflecting the real, highly imbalanced nature of the data, and avoids spending training steps on correcting the output bias.

Furthermore, the model training was designed with efficiency in mind, to bring the training and retraining costs down despite handling large datasets. Thus, gradient accumulation steps were introduced, to split the big batch of 2048 into micro-batches of 512 which would each fit in GPU memory, with the optimizer updating the weights only after all 4 of the micro-batches are computed.

D. Wang et al. (2022) treats position on the page as a special case, essentially joining it with other features only at the very end of the pipeline. The authors argue

that position should be independent of the ad's own qualities. In this thesis, however, I don't see a justification to make the same distinction, so I treat position (pl_position) as just another part of context.

3.5.2 Experiment: dimension reduction

D. Wang et al. (2022) suggests using a dimension reduction step before fusion. Moreover, the paper found a `0.06%` AUC boost (which the authors argue is significant in their context) from using node-wise summation rather than concatenation in the fusion step, which motivates validation of this improvement for our model. Thus, in this thesis two comparisons were carried out: 1) dimension reduction vs. status quo as described above, and 2) node-wise summation vs. concatenation.

Indeed, in the architecture shown in Figure 8 the fusion vector consists of a concatenation of the 768-long `[CLS]` pooling vector from XLM-R and 52 floating-point numbers from non-textual features. This might overweight the impact of the textual tower in predictions. Dimension reduction would also reduce the required size of the head MLP. In this thesis, I implemented dimension reduction as a linear layer on top of the textual tower, as well as one on the non-textual tower. Both project the data on 30 output values. In the node-wise summation case, the resulting fusion vector is of length 30. In the concatenation case, the fusion vector, input to the MLP head, has length of 60.

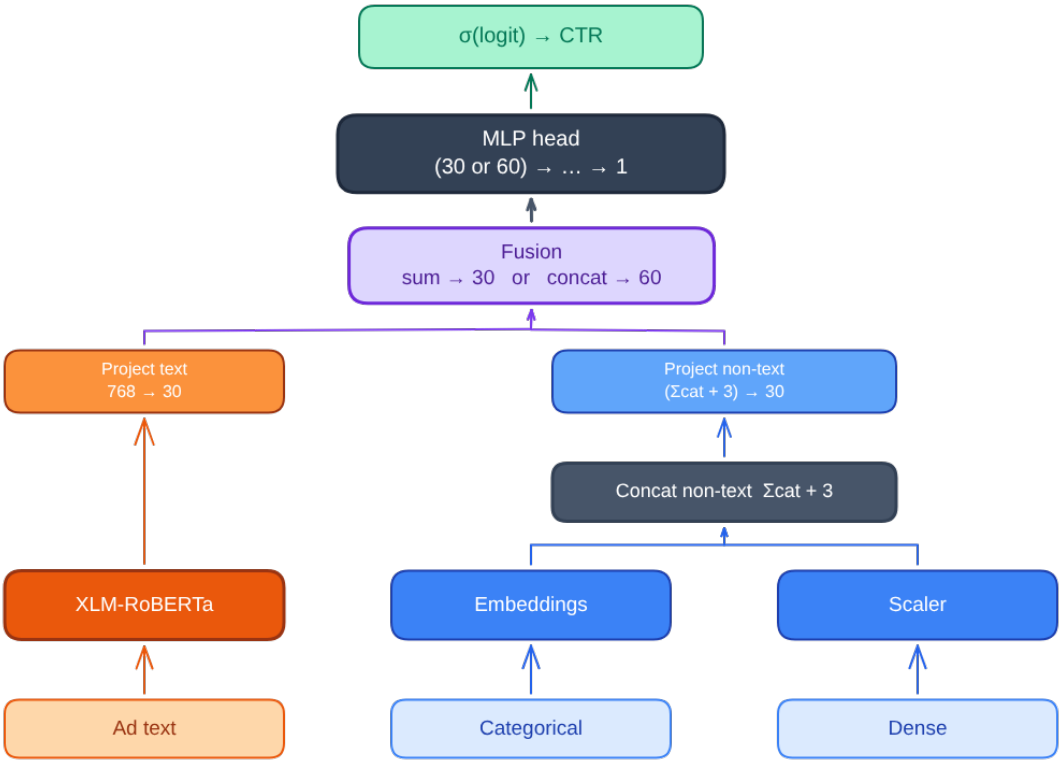

Figure 9: Late fusion with dimension reduction.

Final comparison included three variants: 1) no projection variant as depicted in

Figure 8, 2) linear projections with fusion as concatenation, and 3) linear projections with node-wise summation on fusion step. Since the number of parameters varies significantly (two-layer MLP head in no projection setup; one-layer with different number of inputs for the other two), a separate head learning rate sweep was performed first, to determine the best value for each variant. Other hyperparameters were fixed across these runs, as was the initialization seed. Results of the sweep are shown in Figure 10. The learning rate with the lowest loss was selected for each variant (marked with a star on the figure). 10-seed final comparison of variants, each with its best configuration, is presented in Table 9.

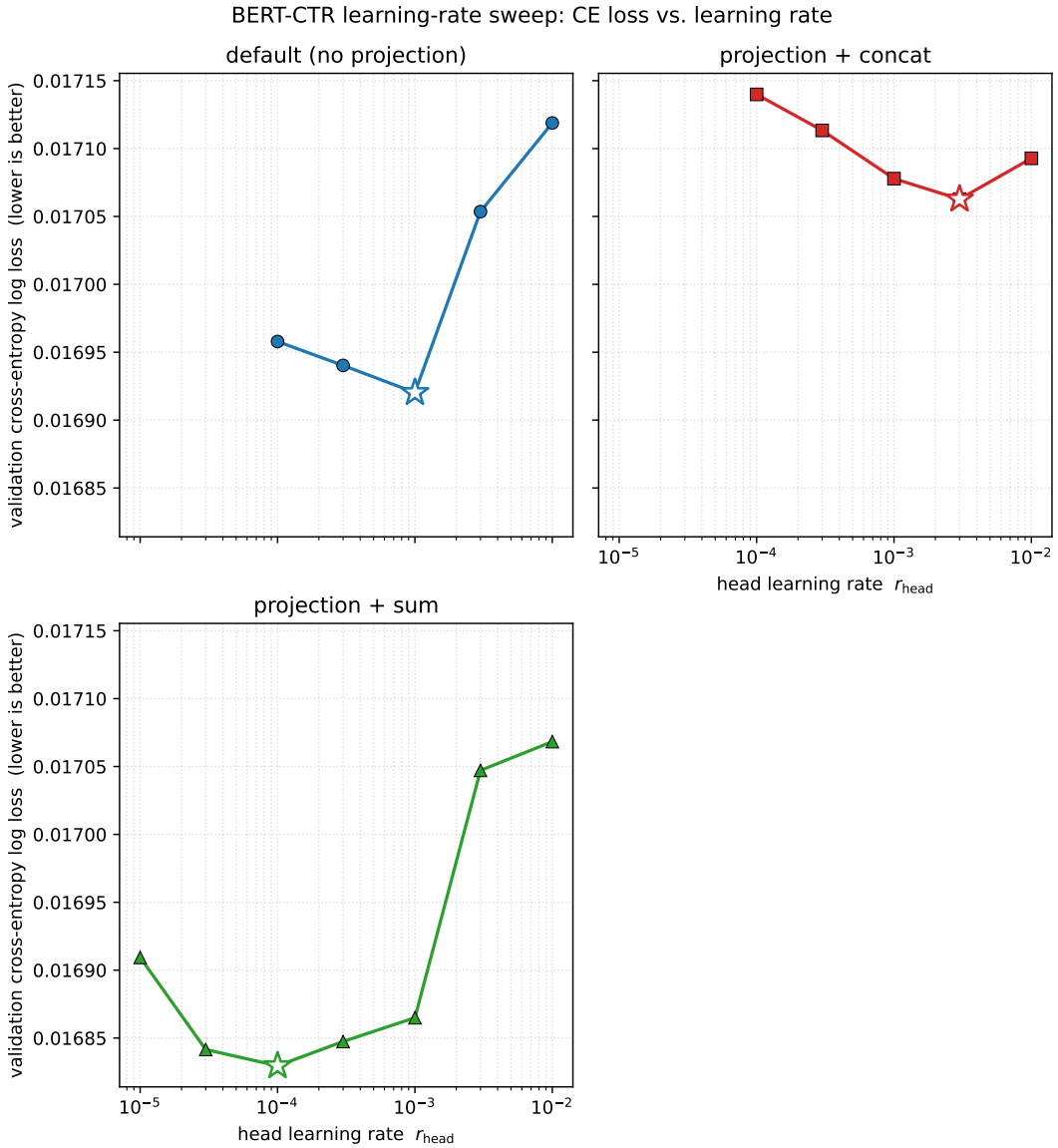

Figure 10: Learning rate sweep over three variants of BERT-CTR architecture.

3.5.3 Implementation

The late fusion architecture has a significant benefit for real-world applications. Since the 270M-parameter transformer only processes the ad and article texts, its output - the 768-dim [CLS] pooling vector - can be precomputed and cached, saving significant time at inference, achieving single-digit millisecond prediction times, even on CPU. Exact numbers are shown in Table 7.

Table 7: Benchmark results (batch size 128, 5000 inference samples / 20 training steps). `train_frozen` denotes training of the model with frozen RoBERTa weights (no fine-tuning). Python code was used, which makes these results an upper bound on cached inference time.

Run	Scenario	Wall (s)	Throughput	Unit	Peak mem (MB)
CPU	infer_full	414.591	12.060	rows/s	-
	infer_cached_cls	0.026	189877.321	rows/s	-
	train_full	826.771	0.024	steps/s	-
	train_frozen	357.251	0.056	steps/s	-
GPU (A100)	infer_full	1.361	3672.885	rows/s	1511.2
	infer_cached_cls	0.018	285359.754	rows/s	1087.2
	train_full	1.835	10.902	steps/s	9777.7
	train_frozen	0.519	38.549	steps/s	1443.5

The benchmark results show that precomputing the `[CLS]` cache makes the real-time computation cost 16'000 times lower, making this architecture attractive for real-world usage. Moreover, the hot path inference may be implemented in a performance-optimized programming language of choice (such as Go), since the MLP, embeddings and sigmoid function contain relatively simple computations. This avoids the traditional overhead of a Python or ONNX environment. In particular, I leveraged these benefits in my implementation for the ad server of C Wire, achieving p99 latency of $50\mu s$, or $0.05ms$ for calls to the prediction service.

Since trends and patterns in advertising have a tendency to change, a DSP implementing such model would have to watch out for model drift. Model drift, or model degradation, is a phenomenon that occurs when the environment is dynamic, and the model's predictive power deteriorates as its training data does not reflect the patterns seen in real-time data. For example, an event like the World Cup might happen, which changes the distribution of football-related content, as well as changes people's predisposition to products around it.

To address the issue of model drift, an adtech operator needs to periodically retrain the model on new data. Luckily, with a late-fusion architecture, the costs of this can be mitigated by freezing the weights of the textual tower for more frequent retrains. For example, the categorical & dense features could be retrained along with the MLP head every week or more often, while the expensive full retrain with fine-tuning the transformer weights could be run every 2-3 months, to incorporate new longer-term trends in data.

4 Results

Table 8: Comparisons of CTR prediction models on validation dataset. Higher AUC, lower log loss and higher RIG are better; the best AUC is shown in bold. Results from initialization sweeps are shown as mean±std

Model	AUC	Log loss	RIG
Baseline (historical CTR)	0.6729	0.01727	–
Logistic regression	0.6840	0.01719	+0.47 %
FFM (best of sweep)	0.6892	0.01702	+1.47 %
BERT-CTR	0.6954 ± 0.0070	0.01690 ± 0.00009	**+2.15 %** ± 0.53 %

Table 8 shows the headline comparison between the models presented in Section 3. Models used the features as noted in Section 3.1, with the following exceptions:

1. The predictions of the baseline (historical averages) were one of the features of all models, so it is expected that all evaluated models beat the baseline.

2. Moreover, the historical averages were computed using data from both training and validation set, instead of freezing them for evaluation. They were computed and evaluated on a rolling basis, where each prediction is calculated only using impressions before the current one. This models the real-world performance the best, but introduces a small degree of validation set leakage to the models, because the historical CTR is one of the input features. Arguably, this evaluates the models in a way closest to their potential behavior when deployed, and it does not affect model vs. model comparisons.

3. BERT-CTR model used textual data from the ad and article (see Section 3.1.2). This naturally adds some performance gain on its own, which is here counted towards BERT-CTR improvement.

4. FFM results might be slightly inflated since only the best result of the hyper-parameter sweep is taken, with no account for possible overfit or initialization dependency, as in BERT-CTR.

Relative information gain (RIG) is shown alongside AUC and log loss to provide an interpretable "lift" value. Note that RIG is computed relative to the baseline model from Section 3.2. More information on definition and calculation of metrics can be found in Section 2.3.4.

Logistic Regression and FFM show small wins over the baseline, with FFM showing 1.47% RIG, and LR only 0.47%. FFM might have shown significantly better results if a collection of high-cardinality features was provided such as those commonly found in personal data.

Transformer-based late fusion model ("BERT-CTR") shows about 2% positive relative information gain and an ≈0.02 AUC lift over the baseline. Alotaibi and Alotaibi (2025) claims that even an increase of 0.001 in AUC is significant when deployed in practice, which makes the results worthwhile.

4.1 Experiment: dimension reduction and node-wise summation

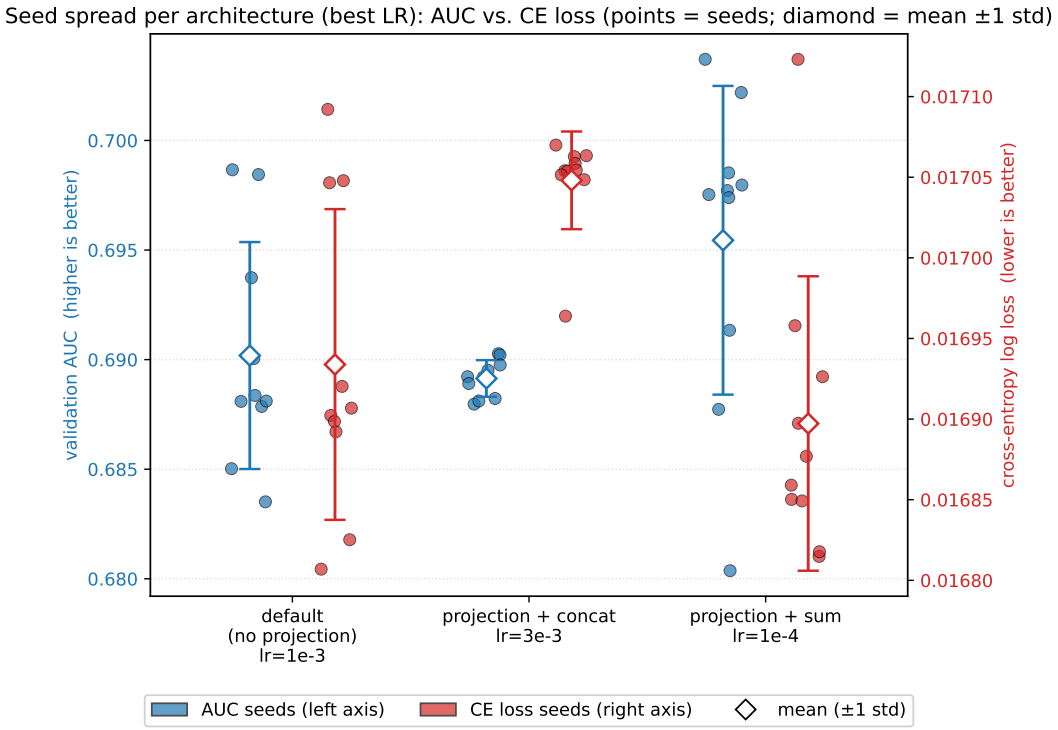

Figure 11: AUC and cross-entropy loss for three variants of BERT-CTR model with different fusion approach, across 10 runs with different initializations.

Table 9: Seed-to-seed variation of BERT-CTR architectures on the validation dataset. AUC and log loss are reported as mean ± sample standard deviation over random seeds; higher AUC is better, lower log loss is better. The last two columns are the ablation CE gains (mean log-loss increase when a tower is removed; higher = that tower contributes more): non-text ΔCE zeroes the categorical + dense towers, text ΔCE zeroes the text tower.

Architecture	AUC	Log loss	Non-text ΔCE	Text ΔCE
default (no projection)	0.6902 ± 0.0052	0.01693 ± 0.00010	0.00069	0.00027
projection + concat	0.6891 ± 0.0008	0.01705 ± 0.00003	0.00079	0.00002
projection + sum	0.6954 ± 0.0070	0.01690 ± 0.00009	0.00061	0.00040

Table 9 and Figure 11 show the results of the experiment described in Section 3.5.2. As can be seen, dimension reduction with node-wise summation ("projection + sum" variant) wins by mean headline metrics. It is notable, however, that this variant also has the highest initialization-related variation in performance, which makes these results more suggestive than statistically significant. At the same time, this variant

results in the lowest parameter count, and the "hot path" computation of the non-textual tower with the MLP head is the lightest compared to other variants.

Additional analysis was carried out to determine the true contribution of the heavy textual tower compared to other signals via ablation. The model was evaluated on the validation set while replacing the output of the non-textual tower with a zero-valued vector, and then separately similarly hiding the output of the textual tower. Increase in the loss in each case was measured and mean value over the runs was recorded in Table 9.

It can be seen that the "projection + concat" variant basically ignores the signal from textual tower altogether. Meanwhile, "projection + sum" gives the textual tower the highest weight across the variants, which seems to result in generally higher AUC and lower loss. Based on this and the above evidence, dimension reduction with node-wise summation is the preferred architecture for this kind of model.

4.2 Calibration analysis

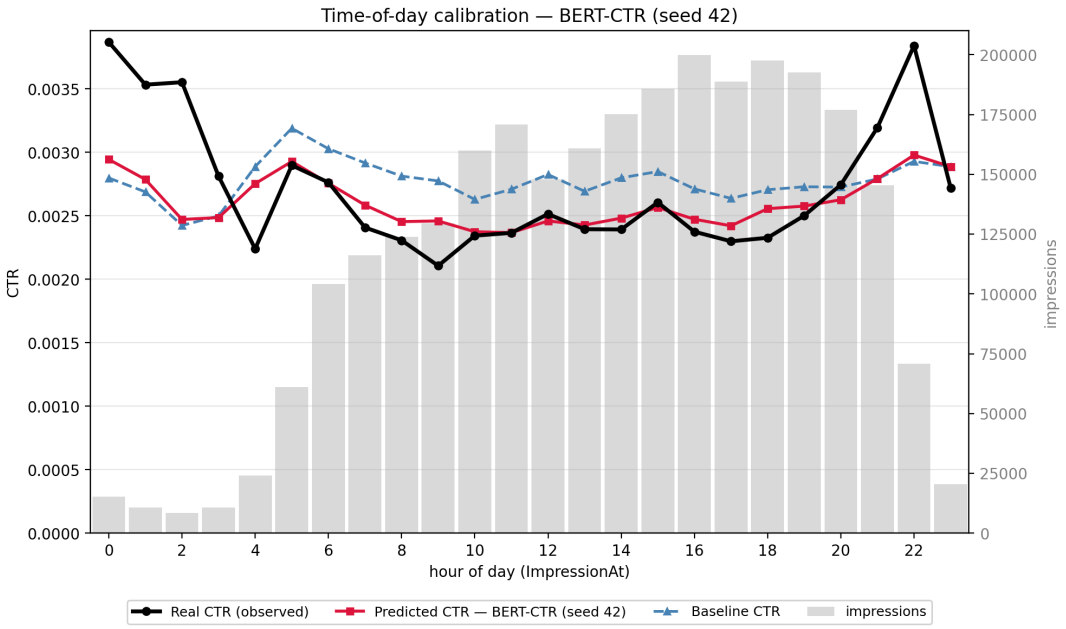

Figure 12: Observed vs. expected CTR by hour of day.

Calibration is an important qualitative metric for CTR prediction. Instead of aggregated, "mean" performance, it shows how the model performs on different subsets of data. Analysis of calibration helps spot potential problems with the model and gaps in its understanding of data. In a real-world implementation, it is crucial to not only predict well on average, but to also avoid very poor performance on specific segments. For example, constant underbidding on an individual site or publisher would result in a possible violation of the contract with them.

Figure 12 shows a plot of calibration by time of day. It can be seen that CTR naturally rises during late evening and night, although there are also significantly fewer

impressions during that time, which might explain why the predicted values do not follow the curve of real CTR as well during that period. There seems to be a dip in real CTR on hour 23, which might be explained by some errors in attributing clicks across days. The baseline predictor seems to be overestimating CTR at most of the hours, compared to the model, predictions of which are more correlated with the real CTR.

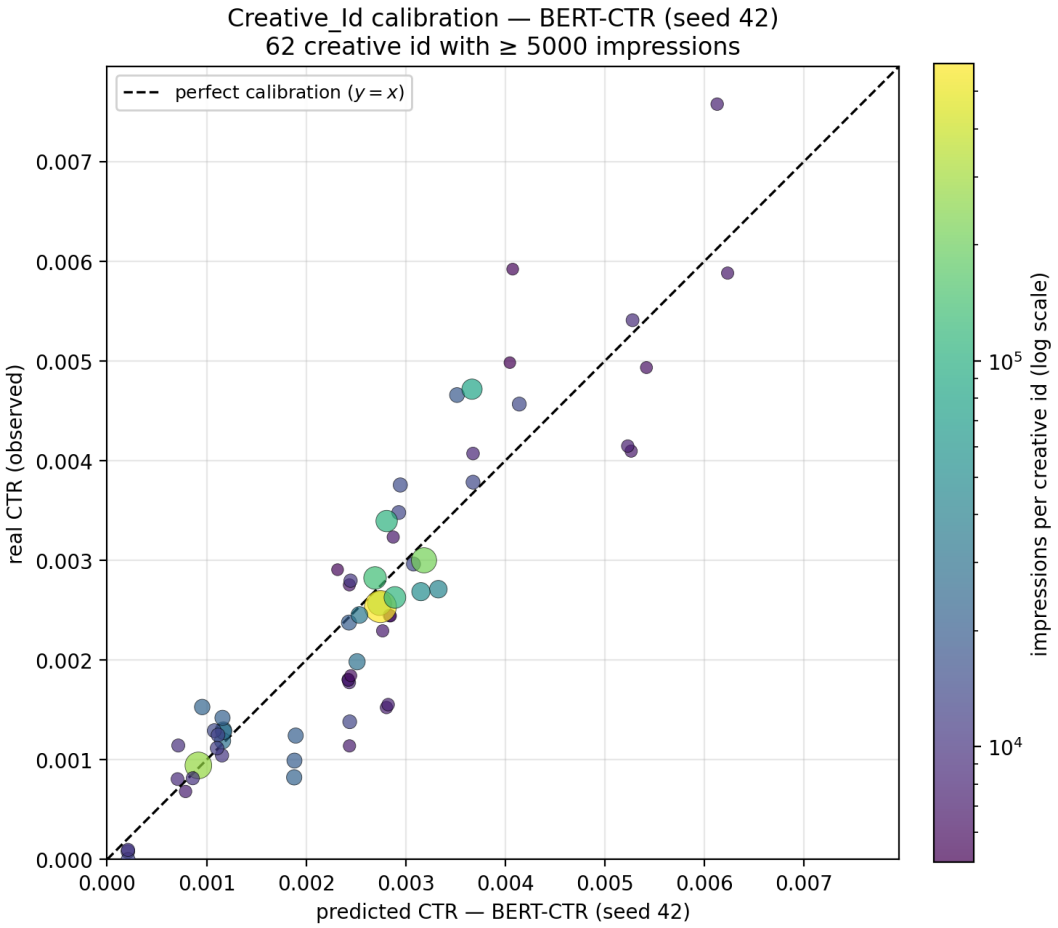

Figure 13: Observed vs. expected CTR by Creative_Id.

Figure 13 shows an observed vs. expected CTR plot, where each point corresponds to a single creative (advertisement). Bigger bubbles correspond to campaigns with more data and traffic. High-traffic campaigns seem to be well-represented by the model (big bubbles are close to the `y=x` line on the plot). A cluster around 0.003 click probability is observed despite some variation in real CTR. This might show that the model could get better performance if it is trained on more data than in this thesis' experiments.

5 Discussion

Evaluation of the models for CTR prediction shows that Transformer-based BERT-CTR model achieves the best performance on the task, compared to the baseline as well as popular algorithms (RQ1). This thesis demonstrates the potential of architectures based on pre-trained language models for CTR prediction and similar tasks in online advertising, especially in the environments where rich personal data is not available.

Results are consistent with findings of D. Wang et al. (2022): node-wise summation and dimension reduction indeed seem to do the best job at capturing the textual and contextual signals (RQ2). The implementation is shown to fit the strict inference requirements, with sub-millisecond latency and good calibration (RQ3).

Textual data is shown to be valuable for predicting click-through rate, as shown using the pre-trained language model RoBERTa, while a Factorization Machine showed decent but worse results. At the same time, the AUC increase observed from BERT-CTR was less than 0.02, with RIG 2%, which both point to a positive yet not overwhelming result. A 2% increase in revenue would be a good result – however, the actual impact on revenue numbers is difficult to predict due to many factors beyond the machine learning task of predicting click-through rate. These include, for example, available ad and placement inventory, and market competition. However, there are also known limitations of the research done in this thesis:

1. Performance of BERT-CTR may be partially explained by introduction of textual fields, unavailable to other models. A more fair comparison could include e.g. TF-IDF embeddings as a feature for FFM.

2. The BERT4CTR paper introduced by D. Wang et al. (2023) points out that late fusion loses out to another approach the authors called Uni-Attention. In that architecture, in addition to final concatenation / summation of tower outputs, the non-textual tower has a set of attention layers that takes Key/Value from textual tower's transformer layers. Since one-directional attention is used, it is still possible to cache the outputs of the textual part to avoid expensive inference (the approach from 3.5.3), however, the Key and Value matrices for each attention layer also need to be cached, raising the amount of memory required per cache entry from kilobytes to megabytes. Uni-Attention was decided to be left out of scope for the current thesis, since basic late fusion offers the proof of concept with lower architectural requirements.

3. Evaluation on real campaigns, for example, via A/B tests, could be carried out to get more realistic measurements of the revenue lift achieved over baseline. As discussed above, the real revenue impact depends on several other factors beyond the CTR estimator itself. Such evaluation decidedly did not fit into the timeline of the current work. However, it is on the roadmap as an essential step in the integration of this model into the production adserver.

4. More deliberation could be done on the exact formulation of what an "ad text" is. In fact, it was more of a product description than text on the ad, since the

information came from the scraped call-to-action (CTA) link of the ad. This enables the information about product-context fit to be modeled by the predictor, but loses the semantics of the ad itself. One could experiment more with the specific way these ad texts are generated.

5. Other events than click could be modeled. Once again, the current thesis offers the proof of concept solution on the most impactful type of event, but a practically identical model could be trained to predict, for example, the probability of a full view or engagement by an arbitrary definition for which there is enough data. This would enable a DSP to offer competitive cost-per-action pricing schemes to its customers.

6. Finally, an inherent limitation is present in the dataset used. J. Wang et al. (2017) highlights that the data about won bids collected by a single DSP is biased, because only the winning price for the auctions that this DSP won is known.

6 Conclusion

This thesis investigated the problem of online ad targeting in the context where personal data is not available. Personal data is becoming more scarce and problematic to handle, while contextual data is now richer and easier to process than ever, due to breakthroughs in the NLP domain. Moreover, contextual targeting may offer better long-term outcomes for all participants of the ecosystem than tracking-enabled personalized targeting.

The paper examines solutions to the cornerstone problem in digital advertising - click-through rate prediction, which is an important part of the optimization of online ad campaigns. Using data from a Swiss adtech provider, a proof-of-concept model is developed and trained. It is observed to fit the acceptability criteria (reliability and strict latency requirements) and beat both the provider's existing solution and versions of popular CTR prediction algorithms, adapted to the no-tracking context. In particular, a two-tower architecture is proposed, where heavy pre-trained language model inference may be pre-computed and cached. The trained model is observed to offer a 2% Relative Information Gain and 0.02 AUC lift, while being well calibrated and showing sub-millisecond times on cached inference.

This thesis is meant to test whether online advertising is possible without collection and processing of granular personal data. Indeed, according to my literature review, studies disagree on the actual added value of tracking, while there is mounting evidence of its negative impact for consumers, advertisers and publishers alike. In this thesis, I go further and show a potential alternative: a model that leverages the depth of available contextual data to deliver advertising outcomes in complete absence of tracking.

Online advertising helps promote new brands and products, and supports webpages that offer free access for all, from news media to personal blogs. Contextual advertising, where the displayed ad is relevant to content around it, is a simple yet powerful method

of ad targeting, with obvious relevance to the reader. Such an approach avoids damage to personal privacy and does not even require e.g. collecting timestamped browsing or location history, or inferences about interests and mental state. As an added benefit, there are vastly reduced risks of abuse and data leaks.

Testing in a production scenario, on running ad campaigns, is to follow the current research. Limitations discussed in the previous section show possible future improvements to the model design. Future research could also focus on experiments with new architectures and variants, as well as more rigorous evaluation of the model introduced here. Moreover, a proper comparison against a personal-data-powered model could be done on live data.

I hope that this thesis contributes to more research of privacy-friendly solutions in advertising technology. While AI offers novel forms of hyperpersonalization, it also unlocks heaps of new value in contextual data, and with it new long-term benefits to consumers and industry.

References

Alotaibi, S., & Alotaibi, B. (2025). A review of click-through rate prediction using deep learning. *Electronics (Basel)*, *14*(18), 3734.

Bai, J., Geng, X., Deng, J., Xia, Z., Jiang, H., Yan, G., & Liang, J. (2025). A comprehensive survey on advertising click-through rate prediction algorithm. *The Knowledge Engineering Review*, *40*, e3. https://doi.org/10.1017/S0269888925000025

Busch, O. (Ed.). (2015, November). *Programmatic advertising 2016* (2016th ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-25023-6

California Department of Justice, Office of the Attorney General. (2024). California consumer privacy act (CCPA) [Accessed: 2026-06-09].

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., & Shah, H. (2016). Wide & deep learning for recommender systems. *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, 7–10. https://doi.org/10.1145/2988450.2988454

Citron, D. K., & Solove, D. J. (2022). Privacy harms. *BUL Rev.*, *102*, 793.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 8440–8451.

Davies, J. (2017, March). *The gloves are off: The Guardian sues Rubicon Project for undisclosed fees*. Digiday. Retrieved June 8, 2026, from https://digiday.com/media/gloves-off-guardian-sues-rubicon-project-undisclosed-fees/

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186.

European Commission. (2025, November). *Digital package* [Last updated 20 November 2025]. Shaping Europe's Digital Future. Retrieved June 12, 2026, from https://digital-strategy.ec.europa.eu/en/faqs/digital-package

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of economic perspectives*, *23*(3), 37–60.

Fouad, I., Bielova, N., Legout, A., & Sarafijanovic-Djukic, N. (2018). Tracking the pixels: Detecting web trackers via analyzing invisible pixels. *CoRR*, *abs/1812.01514*. http://arxiv.org/abs/1812.01514

Gu, Z., Johnson, G. A., & Kobayashi, S. J. (2026). Can privacy technologies replace cookies? ad revenue in a field experiment. *Proceedings of the National Academy of Sciences*, *123*(19), e2603752123. https://doi.org/10.1073/pnas.2603752123

Guo, H., Tang, R., Ye, Y., Li, Z., & He, X. (2017). Deepfm: A factorization-machine based neural network for ctr prediction. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1725–1731.

Häglund, E., & Björklund, J. (2024). Ai-driven contextual advertising: Toward relevant messaging without personal data. *Journal of Current Issues & Research in Advertising*, *45*(3), 301–319. https://doi.org/10.1080/10641734.2024.2334939

Interactive Advertising Bureau & Media Rating Council. (2025, November). *IAB and MRC attention measurement guidelines* (Industry Guidelines No. Final Version 1.0). Interactive Advertising Bureau. Retrieved May 25, 2026, from https://www.iab.com/wp-content/uploads/2025/11/IAB_MRC_Attention_Measurement_Guidelines_November_2025.pdf

Interactive Advertising Bureau & PwC. (2026, April). *IAB/PwC internet advertising revenue report: Full year 2025*. Retrieved June 29, 2026, from https://www.iab.com/insights/internet-advertising-revenue-report-full-year-2025/

Juan, Y., Zhuang, Y., Chin, W.-S., & Lin, C.-J. (2016). Field-aware factorization machines for ctr prediction. *Proceedings of the 10th ACM Conference on Recommender Systems*, 43–50. https://doi.org/10.1145/2959100.2959134

Laub, R., Miller, K. M., & Skiera, B. (2023). The economic value of user tracking for publishers. *arXiv preprint arXiv:2303.10906*.

Liu, D. C., & Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical Programming*, *45*(1), 503–528. https://doi.org/10.1007/BF01589116

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marotta, V., Abhishek, V., & Acquisti, A. (2019). Online tracking and publishers' revenues: An empirical analysis. *Workshop on the Economics of Information Security*, 1–35.

MDN contributors. (2025, October). Using HTTP cookies [Last modified Oct 8, 2025; accessed June 5, 2026]. *Mozilla*. https://developer.mozilla.org/en-US/docs/Web/HTTP/Guides/Cookies

Metzger, Z. (2025). The state of local news 2025 [Accessed: 2026-06-08].

Moradi, P., Cheyre, C., & Acquisti, A. (2025). Economic rationales for regulating behavioral ads. *Available at SSRN 4954788*.

Ou, W., Chen, B., Dai, X., Zhang, W., Liu, W., Tang, R., & Yu, Y. (2023). A survey on bid optimization in real-time bidding display advertising. *ACM Trans. Knowl. Discov. Data*, *18*(3). https://doi.org/10.1145/3628603

Prebid.org. (2026). What is prebid.js? [Accessed: 2026-06-12].

Raza, S., Rahman, M., Kamawal, S., Toroghi, A., Raval, A., Navah, F., & Kazemeini, A. (2026). A comprehensive review of recommender systems: Transitioning from theory to practice. *Computer Science Review*, *59*, 100849. https://doi.org/https://doi.org/10.1016/j.cosrev.2025.100849

Rendle, S. (2010). Factorization machines. *2010 IEEE International Conference on Data Mining*, 995–1000. https://doi.org/10.1109/ICDM.2010.127

Richardson, M., Dominowska, E., & Ragno, R. (2007). Predicting clicks: Estimating the click-through rate for new ads. *Proceedings of the 16th International Conference on World Wide Web*, 521–530. https://doi.org/10.1145/1242572.1242643

Samuel, A., White, G. R., Thomas, R., & Jones, P. (2021). Programmatic advertising: An exegesis of consumer concerns. *Computers in Human Behavior*, *116*, 106657. https://doi.org/https://doi.org/10.1016/j.chb.2020.106657

Sivan-Sevilla, I., Parham, P., & McGuigan, L. (2025). 'cookie-less' identification for/against privacy? *Internet Pol. Rev.*, *14*(3).

Summers, C. A., Smith, R. W., & Reczek, R. W. (2016). An audience of one: Behaviorally targeted ads as implied social labels. *Journal of Consumer Research*, *43*(1), 156–178.

Tiet, T. (2020). *The planning and implementation process of programmatic advertising campaigns in emerging markets.* [Master's thesis, Jyväskylä University School of Business and Economics]. https://urn.fi/URN:NBN:fi:jyu-202005203359

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wadhwa, S., & Bachwani, S. (2025). Knowing me, knowing you: Behavioral tracking and the invisible algorithms that shape our lives. *Journal of Information Technology Teaching Cases*, *0*(0), 20438869251394036. https://doi.org/10.1177/20438869251394036

Wang, D., Salamatian, K., Xia, Y., Deng, W., & Zhang, Q. (2023). Bert4ctr: An efficient framework to combine pre-trained language model with non-textual features for ctr prediction. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5039–5050. https://doi.org/10.1145/3580305.3599780

Wang, D., Yan, S., Xia, Y., Salamatian, K., Deng, W., & Zhang, Q. (2022). Learning supplementary nlp features for ctr prediction in sponsored search. *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4010–4020. https://doi.org/10.1145/3534678.3539064

Wang, J., Zhang, W., & Yuan, S. (2017). Display advertising with real-time bidding (rtb) and behavioural targeting. *Foundations and Trends in Information Retrieval*, *11*(4-5), 297–435. https://doi.org/10.1561/1500000049

Wang, P., Jiang, L., & Yang, J. (2024). The early impact of gdpr compliance on display advertising: The case of an ad publisher. *Journal of Marketing Research*, *61*(1), 70–91. https://doi.org/10.1177/00222437231171848

Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., & Chi, E. (2021). Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. *Proceedings of the Web Conference 2021*, 1785–1797. https://doi.org/10.1145/3442381.3450078

Wolford, B. (2018, November). *What is GDPR, the EU's new data protection law?* GDPR.eu. Retrieved June 9, 2026, from https://gdpr.eu/what-is-gdpr/

Zhang, K., & Katona, Z. (2012). Contextual advertising. *Marketing Science*, *31*(6), 980–994. https://doi.org/10.1287/mksc.1120.0740

Zhang, W., Yuan, S., & Wang, J. (2014). Real-time bidding benchmarking with ipinyou dataset. *CoRR*, *abs/1407.7073*. http://arxiv.org/abs/1407.7073

Zhang, X., Ramachandran, D., Tenney, I., Elazar, Y., & Roth, D. (2020). Do language embeddings capture scales? *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 292–299.