Characterising Topic Familiarity and Query Specificity Using Eye-Tracking Data

Jiaman He
RMIT University
Melbourne, Australia
jiaman.he@student.rmit.edu.au

Zikang Leng
Georgia Institute of Technology
Atlanta, USA
zleng7@gatech.edu

Dana McKay
RMIT University
Melbourne, Australia
dana.mckay@rmit.edu.au

Johanne R. Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Abstract

Eye-tracking data has been shown to correlate with a user's knowledge level and query formulation behaviour. While previous work has focused primarily on eye gaze fixations for attention analysis, often requiring additional contextual information, our study investigates the memory-related cognitive dimension by relying solely on pupil dilation and gaze velocity to infer *users' topic familiarity* and *query specificity* without needing any contextual information. Using eye-tracking data collected via a lab user study ($N = 18$), we achieved a Macro F1 score of 71.25% for predicting topic familiarity with a Gradient Boosting classifier, and a Macro F1 score of 60.54% with a k-nearest neighbours (KNN) classifier for query specificity. Furthermore, we developed a novel annotation guideline – specifically tailored for question answering – to manually classify queries as Specific or Non-specific. This study demonstrates the feasibility of eye-tracking to better understand topic familiarity and query specificity in search.

CCS Concepts

• **Information systems** → **Users and interactive retrieval**.

Keywords

Eye Tracking, Topic Familiarity, Query Specificity

ACM Reference Format:
Jiaman He, Zikang Leng, Dana McKay, Johanne R. Trippas, and Damiano Spina. 2025. Characterising Topic Familiarity and Query Specificity Using Eye-Tracking Data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3726302.3730174

1 Introduction

When people search for information, they are typically driven by their realisation of a knowledge gap [3], which motivates them to recall relevant prior knowledge. Search begins when individuals feel motivated to address that gap. As motivation builds, they start looking for information to bridge that gap, which involves a cognitive process of integrating new with old knowledge.

A user's familiarity with a topic has been defined as the extent of the user's prior knowledge [9]. Understanding this familiarity allows search systems to tailor results to the searcher's knowledge level, which is also linked to the query formulation behaviour[13]. Query specificity is a critical factor in interpreting user intent [19].

By analysing both *topic familiarity* and *query specificity* through physical responses, search systems could adapt to users' needs in real-time. Pupil responses and eye movements reflect cognitive process like memory and recognition [18]. With mobile and wearable devices (e.g., Tobii Pro Glasses 3 [36], Apple Vision Pro [26]) now supporting eye-tracking [29], such data can be captured in real-world scenarios.

While eye-tracking data has been effectively used to characterise how people pay attention to search results in search engine results pages (SERPs) [7, 12, 24, 30, 33, 34], using eye-trackers to characterise the realisation of an information need and the query formulation in search processes remain under-explored [27].

This work aims to characterise *topic familiarity* and *query specificity* using only eye-tracking data, which could inform the development of an adaptive information retrieval (IR) system based on how people look at the screen. Our contributions are as follows:

(1) It is feasible to classify users' knowledge level in real time using pupil dilation and gaze velocity, without relying on additional contextual information, as demonstrated by our *topic familiarity* prediction (Macro F1 = 71.25% with a Gradient Boosting classifier).
(2) We developed a novel methodology to annotate the specificity of users' search queries. Using this method, we categorised a set of 83 queries as either Specific or Non-specific.[1]
(3) It is feasible to classify query formulation behaviour using pupillary data alone as demonstrated by our *query specificity* classification (Specific vs. Non-specific), which achieved a Macro F1 score of 60.54% with a KNN classifier.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730174

[1] The annotated dataset is publicly available at https://github.com/peanutH/Familiarity-QuerySpec, and includes the queries and their specificity classifications

2 Related Work

Topic Familiarity in Information Seeking. Information seeking begins with an understanding that there is something we do not know (an Anomalous State of Knowledge, ASK) and concludes with that knowledge gap being resolved [3]. This process includes integrating prior knowledge and searching for new information, which may reveal other knowledge gaps. Information seeking is thus a cyclical, iterative process, as described by Marchionini [31]. When people notice an information gap, they review their existing knowledge; if their knowledge falls short, they seek additional information. This cycle repeats until the knowledge gap is resolved, whereupon information-seeking stops.

Aligned with this search cycle is the dual-process cognitive model, which posits that memory involves two distinct mechanisms: *recollection* and *familiarity* [1]. Recollection involves retrieving details about past events, while familiarity enables individuals to distinguish between previously encountered and new information [45]. Together, these processes help people recognise familiar topics and identify novel ones. Research has shown that such memory-related processes can be inferred from eye movement patterns [39]. Prior work has also demonstrated that a user's domain knowledge can be inferred from their eye movements [10], suggesting that search tools could leverage real-time eye-tracking data to estimate user knowledge, allowing for more personalised search experiences.

Query Specificity. To form a sentence about an idea, including constructing a query, we first need at least some understanding of the idea [44]. Query formulation thus occurs after cognitive memory retrieval and activation of prior knowledge. As topic familiarity, defined as the extent of prior knowledge [9], has been shown to influence query formulation behaviour [25], query specificity offers an opportunity to interpret user intent [19].

Early work in this space focused on keyword queries, but searchers increasingly use natural language (NL) queries, leading to more diverse query structures and intentions. To address the evolving query forms, we developed an innovative annotation method grounded in NL questions and answers to classify query specificity.

In information-seeking tasks, the meaning of a question lies in its set of possible answers [38]. According to Hamblin [23], understanding a question's meaning requires knowing what qualifies as an appropriate answer. We formulated guidelines for annotating queries as Specific and Non-specific. The methodology for query specificity classification is discussed in Section 3.1.

Eye Features. Previous work has explored eye fixations with query suggestions [13] and the inference of domain knowledge [10]. Our work builds on the idea that topic familiarity and query formulation are memory-related cognitive processes. We investigated eye features associated with human memory. In particular, pupil size has been shown to effectively isolate cognitive effects of familiarity from physical stimulus characteristics [16]. Furthermore, research has shown distinct pupillary responses based on familiarity with branded products, highlighting the link between pupil dynamics and recognition processes [16]. In addition to pupil size, eye gaze has been used to automatically detect internal states of familiarity [8], as eye movement patterns are closely linked to cognitive memory recall [6].

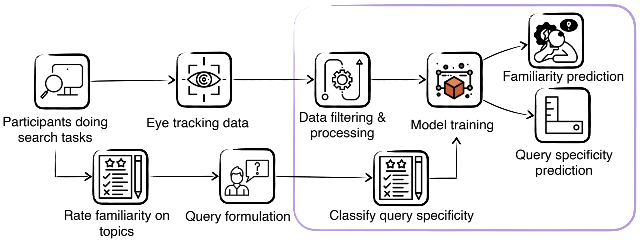
Participants doing search tasks
Eye tracking data
Data filtering & processing
Model training
Familiarity prediction
Query specificity prediction
Rate familiarity on topics
Query formulation
Classify query specificity

Figure 1: Flow of the experiment; our work enclosed in box.

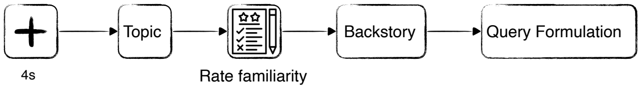

Figure 2: Experiment overview by Ji et al. [27].

3 Experimental Evaluation

Participants performed a search task, during which eye-tracking data, topic familiarity ratings, and query specificity were collected (Section 3.1). Figure 1 illustrates the experimental pipeline. The data was preprocessed and cleaned (Section 3.2) before training five classifiers to predict *topic familiarity* and *query specificity* (Section 3.3).

3.1 Dataset

We used user study data collected by Ji et al. [27]. In the study, participants focused on a cross in the centre of a blank screen for 4 seconds. Then, one of the 12 topic titles appeared, and participants rated the familiarity on a 5-point scale. Next, a backstory was provided for the query formulation, where the backstory was selected from the *InformationNeeds* dataset [2] to evoke the users' realisation of the knowledge gap or information need. The Tobii Pro Fusion eye-tracker[2] was used to collect eye tracking data (60Hz). Figure 2 shows an overview of the procedure.

To ensure the dataset's consistency, we use only written queries in the experiment by Ji et al. [27], as voice input influences query characteristics such as length, terminology, and language [20, 40, 41]. We then analyse the eye-tracking data collected during the phases before the backstory phase, including pupil size and eye gaze.

We re-scaled *familiarity*, measured on a 5-point to binary scale (*1* as unfamiliar and *2* familiar). This approach aligns with familiarity as a signal detection process, where individuals assess whether something feels familiar or unfamiliar [45]. Specifically, 1–3 ratings were grouped as 1 (unfamiliar), while 4–5 as 2 (familiar). We classified a rating of 3 as unfamiliar, based on the assumption that participants selecting this midpoint were uncertain about their familiarity, which we interpreted as indicative of a lack of familiarity.[3]

To annotate query specificity, Hafernik and Jansen [22] showed that queries can be classified as *Specific* or *Non-specific* using a list of attributes. However, their work, conducted over a decade ago, focused solely on keyword-based queries. In contrast, our work examines mostly NL queries, and when three authors applied these

[2] https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-fusion
[3] We have additionally experimented with classifying the rating of 3 as familiar and observed similar model performance trend with overall lower performance.

Table 1: Number of instances categorized by topic familiarity and query specificity .

Topic Familiarity		Query Specificity	
Unfamiliar (49)	Familiar (21)	Non-Specific (60)	Specific (23)

attributes to our dataset, the resulting annotator agreement was low. Next, we developed detailed query specificity annotation guidelines:

(i) Does the query have a clear objective answer? (Yes → jump to *(ii)* ; No/Not sure → Non-specific)

(ii) Does the query require exhaustively listing all valid propositions? (Yes → jump to *(iv)*; No/Not sure → jump to *(iii)*)

(iii) Is it a single/unique answer? (Yes → Specific; No/Not sure → jump to *(v)*)

(iv) Is the answer set bounded? (Yes → Specific; No/Not sure → Non-specific)

(v) Does the answer suffice to be useful and not misleading without requiring exhaustivity? (Yes → Specific; No/Not sure → Non-specific)

For keyword-based queries, we treated them as NL queries by reframing them as questions. For example, the keyword query "greek plays" is re-imagined as "What are Greek plays?" to align with NL query processing.

Three authors independently annotated the queries, with the final annotation determined by majority agreement. We observed a full agreement of 71% and a Fleiss' Kappa [15] of 0.56 among annotators. Table 1 presents the distribution of topic familiarity and query specificity. An example of a Specific query is "What are the sources of slate stone for decorative use, and how is it obtained?" and an example of a Non-Specifc query is "What is the price of the stone?". After preprocessing and cleaning the data (Section 3.2), we retained data from 18 participants, with each contributing between 2 and 6 queries. Note that the total number of data points for topic familiarity and query specificity differs, as not every participant provided a familiarity rating for every topic.

3.2 Data Preprocessing and Cleaning

The baseline data is eye-tracking data from the 4-second cross phase, where participants fixated on a central cross. We first downsampled the 60 Hz eye gaze data to 30 Hz, as recommended for pupillometry studies [43]. Next, we followed the process as described by Franzen et al. [16]. The baseline preprocessing and product-viewing stage pupil data included interpolation of blinks, data smoothing, subtractive baseline correction [32], removal of trials with numerous missing and/or outlier samples, and frequency downsampling. The subtractive baseline correction involved subtracting the median pupil value calculated from the last 150 ms of the baseline stage, just before the transition to the topic page [32]. When pupil data was missing for one eye, we imputed the missing value using the corresponding value from the other eye [4]. The pupil feature was then calculated as the mean of both eyes' values.

Previous research has demonstrated that familiar product images are associated with larger average and peak pupil dilation, with a prolonged response beginning around 1400 ms post-stimulus, indicating differences in familiarity across participants [16]. Based on this, we analysed eye data starting at 1.5 seconds and tested at 300 ms intervals. This approach aligns with findings that real effects on pupil size, such as constriction or dilation in response to stimuli, typically require at least 220 ms to manifest [32].

For the analysis, we computed the Relative Pupil Dilation (RPD), which measures the relative change in the current pupil diameter compared to a baseline value [16, 32]. RPD is defined as: $RPD_t = (P_t - P_{\text{baseline}})/P_{\text{baseline}}$. Here, P_t represents the pupil dilation at time t, and P_{baseline} is the baseline pupil dilation value, calculated as the average pupil dilation over the 4-second cross-phase.

Additionally, we calculated the gaze velocity using the formula: $v = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}/\Delta t$ where (x, y) represents the coordinates of the eye gaze on the screen, and Δt is the time difference between two successive measurements.

3.3 Classifier Training

We use five classifiers to predict users' familiarity with a topic and query specificity based on eye-tracking data: Random Forest [5], KNN [14], Logistic Regression [11], Gradient Boosting [17], and Decision Tree [37]. Deep learning classifiers were not used, as they typically require large datasets to perform effectively.

We computed statistical and signal-based features using the extracted RPD and gaze velocity data, including mean, standard deviation, skewness, kurtosis, zero crossings, and peak-related metrics. To refine the feature set, we applied Recursive Feature Elimination with Cross-Validation (RFECV) [21], which iteratively removed the least informative features. For the topic familiarity prediction, the final selected features included gaze velocity skewness, skewness, kurtosis, root mean square, and the number of peaks for RPD. For the query specificity prediction, only the number of peaks for RPD is used as the feature for the classifier.

For evaluation, we used two approaches: five-fold stratified cross-validation and leave-one-subject-out (LOSO) cross-validation [27, 28], which evaluates the models' ability to generalise to unseen users, thereby ensuring robustness in user-independent scenarios. For five-fold stratified cross-validation, we report the mean and standard deviation of accuracy, macro F1 score, and area under the curve (AUC) across all folds. We report the average macro F1 score across all users for LOSO cross-validation. Additionally, we conducted a paired Student's t-test to compare model performance across 18 folds (1 fold per participant) under the LOSO cross-validation setting. Specifically, we individually compared the best-performing model under LOSO cross-validation against the other four models.

4 Results

Predicting Topic Familiarity. Table 2 presents the performance of various models in predicting users' familiarity with a topic based on eye-tracking data. In stratified cross-validation, the Logistic Regression classifier achieved the highest accuracy (81.81%), while the Gradient Boosting classifier achieved the highest macro F1 score (71.25%), and the KNN classifier obtained the highest AUC (82.61%). Under LOSO cross-validation, the Logistic Regression classifier achieved the highest average macro F1 score (63.74%). To assess statistical significance, we conducted a paired Student's t-test with Bonferroni correction to compare the Logistic Regression classifier's

Table 2: Effectiveness ($\mu \pm \sigma$) of binary classifiers for topic familiarity using eye-tracking data.

Model	Accuracy	Macro F1	AUC	LOSO F1
Logistic Regression	**81.81 ± 7.34**	70.06±16.88	76.27 ± 5.18	**63.74**
KNN	77.52±10.49	71.17±15.65	**82.61±12.84**	50.76
Gradient Boosting	77.81 ± 2.31	**71.25 ± 3.06**	68.64 ± 8.68	55.21
Random Forest	69.52 ± 4.82	57.47 ± 5.13	70.48±14.76	49.69
Decision Tree	72.29 ± 3.84	66.39 ± 5.85	68.55 ± 6.71	51.71

Table 3: Effectiveness ($\mu \pm \sigma$) of binary classifiers for predicting query specificity using eye-tracking data.

Model	Accuracy	Macro F1	AUC	LOSO F1
Logistic Regression	70.09 ± 2.59	48.02 ± 9.86	70.91 ± 6.37	**58.11**
KNN	**71.24 ± 6.44**	**60.54±12.70**	72.86 ± 8.27	53.34
Gradient Boosting	69.90 ± 5.03	44.59 ± 5.21	**74.09 ± 8.30**	44.17
Random Forest	69.90 ± 5.03	44.59 ± 5.21	74.09 ± 8.30	40.31
Decision Tree	69.90 ± 5.03	44.59 ± 5.21	74.09 ± 8.30	44.17

performance in LOSO cross-validation against other models. At $\alpha/4 = 0.0125$, no statistically significant differences were found.

Predicting Query Specificity. Table 3 presents the results for query specificity prediction. In stratified cross-validation, the KNN classifier achieved the highest accuracy (71.24%), and macro F1 score (60.54%). Gradient Boosting outperformed other models in terms of AUC (74.09%). For LOSO cross-validation, the Logistic Regression classifier demonstrated the strongest performance with an average Macro F1 score of 58.11%.

We performed a paired Student's t-test with Bonferroni correction to compare the Logistic Regression classifier's performance in LOSO cross-validation against the other models. Using a significance level of $\alpha/4 = 0.0125$, we did not find statistically significant differences in performance.

5 Discussion

Our result shows the feasibility of using eye-tracking data as a real-time cognitive measurement to predict users' topic familiarity and query specificity. However, we observed a general trend of lower model performance under LOSO cross-validation compared to stratified cross-validation. This discrepancy can indicate that the prediction is more accurate when the model has seen some of a user's eye-tracking data before. Because each person has unique eye movement patterns [35], the model benefits from learning these individual differences, leading to better accuracy when it has access to some of a user's eye-tracking data during training.

For query specificity classification, we used only pupil dilation features. Regarding topic familiarity, four out of five features were also derived from pupil dilation, with only one related to gaze velocity. This suggests that pupillary responses may serve as a stronger indicator than eye gaze in classifying these tasks, likely due to their close relationship with cognitive workload [42]. Additionally, all five models are poorer at predicting query specificity than topic familiarity. This disparity can be attributed to the task timing and the differing cognitive demands they impose. Familiarity ratings were collected after participants read about the topic, aligning closely with the eye-tracking data gathered during the reading phase. This temporal proximity ensured that participants' cognitive state during familiarity rating was consistent with the eye-tracking data.

In contrast, query formulation occurred after familiarity rating and an additional stage involving reading a backstory (Figure 2). These intermediate steps likely altered the participants' cognitive state, making the eye-tracking data collected during the initial topic-reading phase less representative of the mental processes involved in query formulation. Moreover, query formulation is cognitively demanding. It requires translating internal knowledge into structured language, engaging both recognition and recollection-based memory retrieval [44]. This dual memory involvement increases cognitive load and variability, making accurate predictions more difficult.

Topic familiarity, on the other hand, primarily involves recognition—a process associated with lower cognitive effort, as described by the single-detection process model [1]. This makes familiarity easier to infer from eye-tracking data, helping to explain the superior model performance in this task.

6 Conclusions

We explored the potential of eye-tracking data to predict users' topic familiarity and query specificity as a cognitive memory retrieval process. Query specificity was annotated using our newly developed annotation method based on natural language questions and answers. We achieved reasonable performance for familiarity (Macro F1: 71.25%) and query specificity results (Macro F1: 60.54%). These findings suggest that eye-tracking features may offer insights into user needs and goals, which could help inform the design of more personalised and adaptive IR systems.

Limitations and Future Work. In our study, we find it difficult to automatically determine query specificity. So far, we have only analysed the average model performance. Future work could explore features to better distinguish between topic familiarity and query specificity and conduct a more in-depth analysis of query specificity prediction results. Another challenge in this study is the lack of data for models to generalise to unseen users. Additional data could be collected from participants to improve model generalisation. An expanded dataset would also enable exploring advanced models, such as deep learning models, extending beyond this study's five traditional machine learning models.

Acknowledgments

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S, CE200100005), and funded fully by the Australian Government through the Australian Research Council. We thank Kaixin Ji for providing access and support in using the dataset. This work was conducted on the unceded lands of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation. We pay our respect to Ancestors and Elders, past and present, and extend that respect to all Aboriginal and Torres Strait Islander peoples today and their connections to land, sea, sky, and community.

References

[1] Richard C Atkinson and James F Juola. 1972. Search and Decision Processes in Recognition Memory. (1972).

[2] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. Information Needs for TREC 2002-4 (2014). v2. doi:10.4225/08/55D0B6A098248

[3] Nicholas J Belkin. 1980. Anomalous States of Knowledge as A Basis for Information Retrieval. *Canadian journal of information science* 5, 1 (1980), 133–143.

[4] Hal Blumenfeld. 2002. Neuroanatomy Through Clinical Cases.

[5] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. doi:10.1023/A:1010933404324

[6] Andreas Bulling and Daniel Roggen. 2011. Recognition of Visual Memory Recall Processes Using Eye Movement Analysis. In *Proceedings of the 13th international conference on Ubiquitous computing*. 455–464.

[7] Georg Buscher, Andreas Dengel, Ralf Biedert, and Ludger V. Elst. 2012. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Trans. Interact. Intell. Syst.* 1, 2, Article 9 (Jan. 2012), 30 pages. doi:10.1145/2070719.2070722

[8] Iliana Castillon, Trevor Chartier, Videep Venkatesha, Noah S Okada, Asa Davis, Anne M Cleary, and Nathaniel Blanchard. 2024. Automatically Identifying the Human Sense of Familiarity Using Eye Gaze Features. In *International Conference on Human-Computer Interaction*. Springer, 291–310.

[9] John Chesky and Eefrieda H Hiebert. 1987. The Effects of Prior Knowledge and Audience on High School Students' Writing. *The Journal of Educational Research* 80, 5 (1987), 304–313.

[10] Michael J Cole, Jacek Gwizdka, Chang Liu, Nicholas J Belkin, and Xiangmin Zhang. 2013. Inferring User Knowledge Level from Eye Movement Patterns. *Information Processing & Management* 49, 5 (2013), 1075–1091.

[11] David R. Cox. 1958. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)* 20, 2 (July 1958), 215–232. doi:10.1111/j.2517-6161.1958.tb00292.x

[12] Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual differences in gaze patterns for web search. In *Proceedings of the Third Symposium on Information Interaction in Context* (New Brunswick, New Jersey, USA) *(IIiX '10)*. Association for Computing Machinery, New York, NY, USA, 185–194. doi:10.1145/1840784.1840812

[13] Carsten Eickhoff, Sebastian Dungs, and Vu Tran. 2015. An Eye-Tracking Study of Query Reformulation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 13–22. doi:10.1145/2766462.2767703

[14] Evelyn Fix. 1985. *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*. Vol. 1. USAF school of Aviation Medicine.

[15] Joseph L Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological bulletin* 76, 5 (1971), 378.

[16] Léon Franzen, Amanda Cabugao, Bianca Grohmann, Karine Elalouf, and Aaron P Johnson. 2022. Individual pupil size changes as a robust indicator of cognitive familiarity differences. *PloS one* 17, 1 (2022), e0262753.

[17] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189 – 1232. doi:10.1214/aos/1013203451

[18] Stephen D Goldinger and Megan H Papesh. 2012. Pupil dilation reflects the creation and retrieval of memories. *Current directions in psychological science* 21, 2 (2012), 90–95.

[19] Cristina González-Caro, Liliana Calderón-Benavides, Ricardo Baeza-Yates, Libertad Tansini, and Devdatt Dubhashi. 2011. Web Queries: the Tip of the Iceberg of the User's Intent. In *Workshop on User Modeling for Web Applications, WSDM*, Vol. 2011.

[20] Ido Guy. 2018. The Characteristics of Voice Search: Comparing Spoken with Typed-in Mobile Web Search Queries. *ACM Trans. Inf. Syst.* 36, 3, Article 30 (March 2018), 28 pages. doi:10.1145/3182163

[21] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning* 46, 1-3 (2002), 389–422.

[22] Carolyn Theresa Hafernik and Bernard J Jansen. 2013. Understanding the Specificity of Web Search Queries. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1827–1832.

[23] Charles L Hamblin. 1976. Questions in Montague English. In *Montague grammar*. Elsevier, 247–259.

[24] Christopher G. Harris. 2019. Detecting cognitive bias in a relevance assessment task using an eye tracker. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 36, 5 pages. doi:10.1145/3314111.3319824

[25] Rong Hu, Kun Lu, and Soohyung Joo. 2013. Effects of Topic Familiarity and Search Skills on Query Reformulation Behavior. *Proceedings of the American Society for Information Science and Technology* 50, 1 (2013), 1–9.

[26] Apple Inc. 2025. Apple Vision Pro. https://www.apple.com/apple-vision-pro/ Accessed: 2025-01-17.

[27] Kaixin Ji, Danula Hettiachchi, Flora D. Salim, Falk Scholer, and Damiano Spina. 2024. Characterizing Information Seeking Processes with Multiple Physiological Signals. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 1006–1017. doi:10.1145/3626772.3657793

[28] Kaixin Ji, Damiano Spina, Danula Hettiachchi, Flora Dilys Salim, and Falk Scholer. 2023. Examining the Impact of Uncontrolled Variables on Physiological Signals in User Studies for Information Processing Activities. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1971–1975. doi:10.1145/3539618.3591981

[29] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2176–2184.

[30] Jiaxin Mao, Yiqun Liu, Huanbo Luan, Min Zhang, Shaoping Ma, Hengliang Luo, and Yuntao Zhang. 2017. Understanding and Predicting Usefulness Judgment in Web Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 1169–1172. doi:10.1145/3077136.3080750

[31] Gary Marchionini. 1995. *Information Seeking in Electronic Environments*. Cambridge university press.

[32] Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. 2018. Safe and Sensible Preprocessing and Baseline Correction of Pupil-size Data. *Behavior research methods* 50 (2018), 94–106.

[33] Tristan Miller and Stefan Agne. 2005. Attention-based information retrieval using eye tracker data. In *Proceedings of the 3rd International Conference on Knowledge Capture* (Banff, Alberta, Canada) *(K-CAP '05)*. Association for Computing Machinery, New York, NY, USA, 209–210. doi:10.1145/1088622.1088672

[34] Srishti Palani, Adam Fourney, Shane Williams, Kevin Larson, Irina Spiridonova, and Meredith Ringel Morris. 2020. An Eye Tracking Study of Web Search by People With and Without Dyslexia. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 729–738. doi:10.1145/3397271.3401103

[35] Matthew F Peterson and Miguel P Eckstein. 2013. Individual Differences in Eye Movements During Face Identification Reflect Observer-specific Optimal Points of Fixation. *Psychological science* 24, 7 (2013), 1216–1225.

[36] Tobii Pro. 2025. Tobii Pro Glasses 3. https://www.tobii.com/products/eye-trackers/wearables/tobii-pro-glasses-3 Accessed: 2025-01-17.

[37] J. Ross Quinlan. 1986. Induction of Decision Trees. *Machine Learning* 1, 1 (1986), 81–106. doi:10.1023/A:1022643204877

[38] Robert Rooy. 2004. Utility of Mention-some Questions. *Research on Language and Computation* 2, 3 (2004), 401–416.

[39] Jennifer D Ryan, Robert R Althoff, Stephen Whitlow, and Neal J Cohen. 2000. Amnesia is a deficit in relational memory. *Psychological science* 11, 6 (2000), 454–461.

[40] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Information Processing & Management* 57, 2 (2020), 102162. doi:10.1016/j.ipm.2019.102162

[41] Alexandra Vtyurina, Charles L. A. Clarke, Edith Law, Johanne R. Trippas, and Horatiu Bota. 2020. A Mixed-Method Analysis of Text and Audio Search Interfaces with Varying Task Complexity. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval* (Virtual Event, Norway) *(ICTIR '20)*. Association for Computing Machinery, New York, NY, USA, 61–68. doi:10.1145/3409256.3409822

[42] Pavel Weber, Franca Rupprecht, Stefan Wiesen, Bernd Hamann, and Achim Ebert. 2021. Assessing Cognitive Load via Pupillometry. In *Advances in Artificial Intelligence and Applied Cognitive Computing: Proceedings from ICAI'20 and ACC'20*. Springer, 1087–1096.

[43] Matthew B Winn, Dorothea Wendt, Thomas Koelewijn, and Stefanie E Kuchinsky. 2018. Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want To Get Started. *Trends in hearing* 22 (2018), 2331216518800869.

[44] Wilhelm Wundt. 1970. The Psychology of the Sentence. *AL Blumenthal (Ed. and Trans.), Language and psychology: Historical aspects of psycholinguistics* (1970), 20–31.

[45] Andrew P Yonelinas. 2001. Components of Episodic Memory: The Contribution of Recollection and Familiarity. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 356, 1413 (2001), 1363–1374.